

Fading pan-European visions?

Making the most of European collaboration would seem to be sensible as pressures on research budgets intensify. But those with the power to pursue that agenda are failing to do so.

About forty years ago, a grand moment arrived for those visionaries who, since 1945, had been working to increase economic cooperation between European states: the establishment of the common market and the European Economic Community, signified in the Treaty of Rome of 1957. In the post-war years, the idea that science should also be supported in a pan-European manner was almost taken for granted. But, at the time of the treaty, Europe was more preoccupied with a growing technology gap and the 'brain drain' to the United States, and the main focus of attention was on technology-oriented research projects in the aerospace, space and nuclear fields.

That was a shame for fundamental science. For the treaty incorporated the creation of the European Commission, based in Brussels. And for all the charges that have been laid against it, the commission is unique and valuable in at least one respect: it contains a group of bright people mandated by governments to think supranationally, that is, to seek out and nurture the added value that can come out of pan-European collaboration.

At the outset, science was unfortunately not on their agenda. Indeed, to many scientists, the word 'Europe' is a professional irrelevance. After all, ideas and techniques know no geographical boundaries, and so long as researchers receive funds and establish good collaborations, what else matters? But even putting aside the political goal of supporting weaker regions in their scientific ambitions, and concentrating on the pursuit of science at its very best, the benefits of European collaboration need be no greater than that with the United States to make it worth pursuing.

Enthusiasm

That much is clear from the enthusiasm with which researchers apply for the 'training and mobility' segment of the European Union (EU)'s Framework programme in research and development — the only part in which, in the pursuit of encouraging the mobility of European researchers, the commission is allowed to support basic research, even on topics of no immediate economic value. So much is also clear from the positive perspectives and contacts that are formed as evaluators and peer reviewers assess EU research programmes.

A new player in European science politics is the European Parliament which, following the signing of the Maastricht Treaty, now has the capacity substantially to influence scientific policy — as it achieved last week in the scrutiny of the European Commission's proposed fifth Framework programme (see page 3). But the parliament's priorities appear to be the scrutiny of strategic research — and from many perspectives, reflecting the interests of individual politicians and the representations of lobby groups, some of them highly critical of science.

A look elsewhere around Europe today suggests that, although the spirit of Europe's science community is willing, the collaboratively minded flesh of its paymasters is not. No funding institution has both a confident mandate and the clout required successfully to pursue the full added value of pan-European science.

The landscape of facilities highlights the trend. Existing facilities are thriving, but what of the future? CERN, one of the triumphs of scientific collaboration in Europe, is turning into a world facility — its European identity will now be progressively buried under the realities of

maintaining Japanese and US funding — and why not, if that is the natural condition of high-energy physics? The Institut Laue-Langevin and the European Synchrotron Radiation Facility (ESRF) in Grenoble are both performing well. But plans for new synchrotrons in particular are developing at national or *ad hoc* multilateral levels.

As pressures on funding intensify, the value of collaboration and complementarity — in skills, in techniques, in the sharing of large and medium-sized facilities — across Europe should be growing. Yet, in the absence of a substantial role for the commission in fundamental science, what are other paymasters doing about it? An unofficial grouping of European heads of research councils (Eurohorcs) meets every six months to discuss matters of common concern, such as scientific fraud. These individuals control more than 90 per cent of the scientific funding in Europe. Yet they distance themselves from any attempt actively to coordinate or develop activities at a pan-European level.

Hope

The final existing hope for Euro-enthusiasts is the European Science Foundation (ESF), a quasi-independent body funded by national research agencies. Here, if anywhere, is where the pan-European agenda should be being pursued. And to an extent it is. Small collaborative networks and programmes have been established in life sciences, physical sciences and social sciences. 'Euroconferences' are organized jointly with the commission. But the ESF remains a small group of people, located relatively inaccessibly in Strasbourg.

Furthermore, the ESF's role is undermined by the ambivalence of its paymasters, those very bodies represented by the Eurohorcs, about what that role should be. Take, for example, large facilities. Occasionally the foundation is given a key role, as it was in acting as a midwife for the ESRF during the critical phases of its design, and as it is in the current development of proposals for a European Spallation Source (for neutrons) and 100-tesla magnetic field facility. But individual states meanwhile proceed with new neutron sources with scant reference to wider European considerations.

Yet national research agencies have mixed feelings about the ESF taking on a more systematic role in such areas, while the foundation itself is nervous about upsetting them. This confusion as to whether the national agencies are enlightened sponsors, or paymasters in search of direct returns, is a crucial problem for the ESF, which has also failed to position itself as a body that represents Europe's research scientists.

Is the relative lack of support for pan-European research a short-term problem? If one assumes that the ESF is the only body capable of pursuing such an agenda, then the role of its senior figures in maintaining credibility and a reputation for quality is crucial. That is a challenge for the ESF's incoming director-general, Enric Banda. An enlightened goal for the foundation's paymasters would be to build its formal role in making the best of European complementarity, and to give it the mandate to speak out on the obstacles and opportunities facing Europe's researchers. As national agencies, facing budgetary squeezes, seek to retain maximum control over their budgets, the foundation faces an uphill battle in carving a role for itself. But it represents the only opportunity for those trying to make the best of Europe's basic science. □



UK insurers oppose moratorium plea on use of genetic data

[LONDON] Britain's insurance industry appears to be on a collision course with the government after the industry rejected a proposed two-year moratorium on the use of genetic information in assessing applications for life insurance.

The government's Human Genetics Advisory Commission (HGAC) argued in a report just before Christmas that "substantial research" was needed on the actuarial implications of genetics before estimates of health and lifespan could be drawn from genetic tests.

In a policy statement issued a day after the report, the Association of British Insurers (ABI) said that its 440 member companies continued to oppose a moratorium.

The association's statement emphasized that the ABI's code of practice states that the results of genetic tests are not needed for most types of health, medical and critical illness insurance, and for life insurance cover below £100,000 (US\$165,000) linked to



house purchase loans.

But a former member of the House of Commons Select Committee on Science and Technology, whose report on human genetics two years ago led to the creation of the HGAC (see *Nature* 376, 202; 1995),

promised to press for legislation if the call for a voluntary moratorium was ignored.

"The insurance industry is one of the UK's important industries, and we don't want to damage it," says the MP. "But we also want to protect the rights of people. If someone becomes uninsurable, they'll have difficulty in certain fields of work, and in buying a house." In Britain, life insurance is a requirement for some forms of house purchase loan.

The House of Commons science committee gave the insurance industry one year from the publication of its report to come up with proposals on genetic tests, and the HGAC decided to study the issue at its first meeting in February 1997.

Sir Colin Campbell, who chairs the HGAC and is vice-chancellor of the University of Nottingham, says it is not safe to make predictive statements about genetic tests. "It is far too early to be able to reach any conclusions about how genetic testing can be used to predict life expectancy or the onset of ill-health," he says. "Genetic testing is in its infancy."

But Vic Rance, a spokesman for the ABI, says the industry is aware that no genetic test can predict lifespan for most disorders such as heart disease and certain cancers that are caused by a mixture of genes and the environment.

Insurance companies oppose a moratorium as they believe it would be the first step towards a complete ban.

But a spokeswoman for the HGAC says a moratorium is not the same as a ban. "The commission has not at all closed the door on the insurance industry taking in genetic information. They're saying that, in this country, it's too soon."

Insurance companies consider genetic information to be no different from family history information, which is a condition for some policies. The industry says nondisclosure of genetic test results would enable people with genetic illnesses to take out large insurance policies without having to tell the truth about their health. This is known as 'adverse selection'.

They also argue that disclosure has benefits for the insured. A negative test for a person with a family history of a disease will enable that person to obtain insurance cover at standard rather than higher rates.

But the commission's report says the industry can withstand limited adverse selection. It also says there is "widespread concern" that people will refuse to take genetic tests if insurance companies are allowed to see the results.

Agreement likely on Framework budget

[MUNICH] After two days of intense negotiations, the European Parliament last month finally approved the European Union's fifth Framework research programme (FP5), due to start next year. The budget will be ECU16.7 billion (US\$18.6 billion) — only slightly higher than the ECU16.3 billion put forward by the European Commission.

As a result, a prolonged battle over the FP5 budget is now unlikely with the council of ministers, which represents the member states and is keen to keep funding as low as possible. But more problems are likely to arise over the structure of the programme.

To simplify the management of FP5 and concentrate resources, the commission wants to group projects into three thematic programmes: the living world and ecosystems, information and communication technologies, and sustainable growth. These would be accompanied by three 'horizontal programmes' for international cooperation, participation of small and medium-size enterprises, and a training programme called improving human potential.

The council has yet to take a formal position, but has indicated its desire to see five programmes, separating energy and environment from life sciences and giving transport a programme of its own. In contrast, the parliament does not want a

special transport programme — but it does want energy and environment to be separated from life sciences into a single programme in two parts, each with its own defined budget line.

Research commissioner Edith Cresson is likely to fight this latter division, concerned that a separate environmental research programme would be vulnerable to a takeover bid from the environment commissioner.

The parliament also wants to change the commission's proposed distribution of the FP5 budget, voting to restrict the money going to nuclear research to ECU1.3 billion, compared to the commission's proposed ECU1.5 billion.

Parliament does not have decision-making responsibility for this part of the programme, which is covered by the Euratom treaty. And differences of opinion between the two major political groups on support for nuclear fusion — the socialists would like to see much less support than the Christian democrats — meant that no position was taken on how this nuclear research budget should be divided.

The council of research ministers will meet on 12 February to decide its own position, which must be agreed unanimously. If all goes well, calls for research proposals could be put out before the end of the year.

Alison Abbott

Ehsan Masood

CERN celebrates end to a period of uncertainty

[MUNICH] Amid a spirit of unusual cheer, the council of the Geneva-based European Laboratory for Particle Physics (CERN) last month elected Luciano Maiani as its next director-general from 1999. Maiani is head of Italy's Institute of Nuclear Physics, the national research council for particle physics, and head of the CERN council.

CERN has managed to shed the gloom that overshadowed its Christmas last year when member states voted to cut its budget, thereby threatening the future of the planned Large Hadron Collider (LHC), a multi-billion dollar project (see *Nature* 384, 203; 1996).

Since then, the present director-general of CERN, Christopher Llewellyn Smith, has signed several agreements with non-member states, notably with the United States and Japan. These agreements are geared to enable the LHC to be constructed on time.

In addition, a cost-saving plan to offer CERN staff extra holidays in lieu of pay has proved an unexpected success. More than 1,500 staff have taken advantage of the scheme, and the savings are being converted into 37 new positions for young scientists.

The CERN council also agreed to extend the life of the Large Electron-Positron Collider (LEP) programme for one year beyond 1999, provided that SFr50 million (US\$35 million) can be raised from outside sources. Maiani, who is a professor of theoretical physics at La Sapienza University in Rome, says that the scientific case for LEP is now very strong, and as director-general he will make it a priority to find the money.

His first priority, he says, will be to ensure the smooth construction of the LHC. "But it is important to ensure that the scientific life of CERN continues in the period of LHC construction," he says. The extension to LEP is an important part of this.

He will also try to develop other experiments with outside resources, such as the proposed neutrino beam which would run from CERN to southern Italy. The experiment's goal is to determine whether neutrinos have mass.

To accept the post of director-general, Maiani has had to step down as head of the CERN council. This post has been taken by Hans Eschelbacher, an experimental nuclear physicist who has worked for many years in German science policy.



New man: Maiani takes over in 1999

Alison Abbott

Court affirms judges' right to reject 'junk science'

[WASHINGTON] The US Supreme Court last month unanimously affirmed the right of trial judges to disallow scientific testimony they believe to be flawed. The decision may cut down on the use of so-called 'junk science' in the courtroom.

A decision handed down on 15 December reversed a ruling by a federal appeals court that a lower court judge had acted improperly in disallowing expert testimony linking polychlorinated biphenyls (PCBs) to cancer.

In the original case, an individual, Robert Joiner, sued General Electric and two other manufacturers of PCBs, claiming they had caused his lung cancer. But the judge threw out testimony from experts on Joiner's side, arguing that the epidemiological and animal studies on which the experts relied had failed to make a convincing connection between PCB exposure and Joiner's particular cancer.

In backing the original judge, the Supreme Court said that any judge's decision to admit or exclude expert scientific evidence should stand, unless it constituted an "abuse of discretion" or gross error of judgement. The court also said that the original judge was right to throw out the Joiner studies.

Legal experts say that the Supreme Court's position is important because it clears up an area of ambiguity from a 1993 decision on scientific evidence. Until now, judges were generally expected to admit any evidence based on legitimate scientific

methodology — even if its conclusions seemed unbelievable or if the methodology was not well suited to the question.

Last month's decision appears to give judges freedom to consider both methodology and conclusions, just as a scientist might in judging the quality of research.

The court's willingness to take on this issue surprised some legal experts, and makes the case "more important for scientists than I thought it was going to be", says Bert Black, an attorney with the Dallas firm Hughes and Luce. Black co-chairs the National Conference of Lawyers and Scientists, set up in 1974 by the American Association for the Advancement of Science (AAAS) and the American Bar Association.

Courts rely increasingly on expert scientific testimony, particularly in cases involving product liability and the health impact of toxic substances. Black says that not allowing judges to evaluate a scientific study's conclusions would lower the quality of their rulings.

Judges may now have to do more homework in areas of scientific complexity. In a separate concurrence to the Supreme Court's opinion, Justice Stephen Breyer wrote that judges should routinely seek the advice of scientists through pre-trial hearings, expert witnesses and specially-trained law clerks. Black says the AAAS is trying to set up a clearing-house of scientific experts willing to advise courts.

Tony Reichhardt

Council resigns as ANZAAS wind-up fails

[SYDNEY] The entire council of the Australian and New Zealand Association for the Advancement of Science (ANZAAS) has resigned after its recommendation to dissolve the organization failed to win sufficient support from members.

Membership of the 109-year-old ANZAAS numbered thousands in the 1980s but is now down to about 400. Annual congresses, once its main source of revenue, have attracted few paying delegates in recent years (see *Nature* 389, 532; 1997).

But a move to wind up the organization was opposed at a heavily attended special meeting, held in eight main Australian cities before Christmas and linked by phone.

Motions presented by the state divisions of New South Wales and Tasmania that ANZAAS should continue, but with its focus shifting from national to regional activities, were supported by simple majorities. The votes, however, were not binding on council members, who argued that there was no alternative to winding up the association.

Bruce McKellar, who chairs the association, then put two resolutions legally required to wind up the organization. One was to approve dissolution, the other to transfer assets of A\$100,000 (US\$65,000) and some functions to a foundation run by the Australian Academy of Science.

This move was supported by ANZAAS president Sir Gustav Nossal, the former director of the Walter and Eliza Hall Institute of Medical Research in Melbourne. But although both motions secured simple majorities, each failed to achieve the three-quarters of the votes required under ANZAAS's constitution.

The departure of council members means a reshaping of ANZAAS may now take place. But some state divisions which voted against carrying on are likely to drop out. These include Victoria where McKellar, a theoretical physicist, resigned as dean of science in the University of Melbourne after a recent external review was highly critical of his administration.

Peter Pockley

Universities face global challenge



Whatever else 1998 brings, one factor is certain: universities across the world will come under increasing strain as

they attempt to meet diverse and often conflicting demands. In the process, those responsible for university research will face increasingly difficult decisions as they seek to satisfy their paymasters.

Many of the tensions reflect the central role universities are taking on as the 'knowledge producers' of modern economies. There is, for example, conflict between governments' desire to see the production of large numbers of highly trained graduates and their growing reluctance to cover the full costs of doing so from the public purse. At the same time, pressure to mould university research to pay its way by producing useful results now, while beneficial in the short-term, may jeopardize its ability to perform effectively in the longer term.

In this survey, *Nature's* correspondents from around the world describe how these tensions are being felt in university research. The clash between traditional aspirations and harsh economic realities is felt perhaps most sharply in Germany, where recent protests have highlighted the urgency of reforms.

In France, the policy of the new government is to place university research back at the heart of national strategy, a move that is creating tensions with other research institutions.

Japan, like Italy, faces the challenge of building a creative, internationally oriented research base in universities while ensuring that this meets the needs of the wider economy. Britain, which has already made some of the structural changes needed to achieve this, is now struggling to maintain its traditional academic ideals under sustained fiscal and political pressure.

Perhaps each country has something to learn from the United States, where a combination of institutional flexibility, a commitment to scientific excellence and a political recognition of the value of knowledge has produced a system that remains far healthier than some are prepared to admit.

There is no US blueprint that can be applied unchanged elsewhere. Each country must, in the years ahead, find ways to adapt these factors to its own situation. For many, 1998 will be judged by their success or failure in doing so.

Post-war ideals trip over modern reality

Alison Abbott

Pressures are mounting to bring in radical reforms to German universities to reduce bureaucracy and give a boost to scientific research.

There is widespread agreement in Germany that the university system has gone badly wrong, dragged down by the very post-war ideals on which it is based: the decentralization of political responsibility for cultural and educational matters, and the guarantee of free places for all high-school graduates. This discontent has been symbolized by the broad support for this autumn's student strikes.

Student numbers have increased more than threefold since 1970. But this massive growth has not been matched by an equivalent increase in academic staff numbers, nor in university financing. Staff numbers have less than doubled, while total university investment as a proportion of gross domestic product has been falling steadily since 1975.

University research has, unsurprisingly, suffered in the process. Professors have to meet increasing teaching commitments at the cost of their research activities, and also face the problem of ageing laboratory equipment. One of the few direct responsibilities of the federal government in this area is the cofinancing of university buildings and large equipment. Yet it is increasingly reluctant to pay towards items that its science advisory body, the Wissenschaftsrat, says are essential if standards of research are to be maintained (see *Nature* 377, 278; 1995).

Dagmar Schipanski, president of the Wissenschaftsrat, says that only a quarter of the recommended building projects — and fewer than half of the recommended purchases of research equipment — will be financed this year. Construction of urgently needed facilities, such as the 'Biologikum' in Munich, a major complex of laboratories and lecture theatres, has been put on hold indefinitely.

Chronic underfunding has led to a widespread perception that the quality of university research is lower than it should be — and to concern that universities are losing their attractiveness as a location for top-class scientists. In fact, it is difficult to tell whether the overall quality of research really is lower than expected. There is no central group to analyse the quantity and quality of research output, and few universities evaluate their own research performances.

It is clear, however, that universities still house some of the best of German research. Recent statistics on research performance in



On the march: Germany's students are not alone in criticizing the state of the university system.

Germany, published in the autumn by *Laborjournal*, a magazine for biomedical scientists, places Germany in third place for virology and fourth place for immunology in a ranking of publication quality worldwide. Within Germany — at least in immunology and virology — university research groups did proportionately as well as non-university research institutes on the basis of their respective expenditure.

But if university research quality really is under threat, deep flaws in the academic structure of universities may be as much to blame as underfunding. It is widely recognized, for example, that professors wield too much power, with jobs for life whatever their research performance, and that young scientists have too few opportunities to gain independent research experience.

Michael Daxner, president of the University of Oldenburg and an expert on the German university system, is one of many who believe that more competition is needed to improve the quality of research. University chairs come with guaranteed research money for which they do not have to com-

pete. But falling institutional funding has already started to force researchers to become more competitive.

Gerhard Neuweiler, professor of zoology at Munich's Ludwig Maximilian University and former head of the Wissenschaftsrat, has long fought for changes that would support bright young scientists while introducing competition for internal university research funds. He points to the new Institute for Experimental Geochemistry and Geophysics of the University of Bayreuth as "a wonderful example of how things could be".

This institute has virtually no academic hierarchy, and full professors have to compete for internal research funds with postdoctoral fellows who run their own independent research projects. A major visitors' programme allows scientists to spend periods ranging from a day to a year at the institute. "Research thrives in this atmosphere," says Neuweiler. "Every university should have something like this."

As far as the financial picture is concerned, individual *Länder* (states) differ in the way in which they are coping with the problems. Richer *Länder* in the south, such as Bavaria and Baden-Württemberg, have been able to ride the financial storm relatively comfortably. Bavaria has even managed to advance the federal government's share of money for urgently needed equipment, hoping that it will be paid back later.

Those in the poorer north have done less well. The University of Hamburg is shedding 50 per cent of its academic staff. And Berlin is to lose two-fifths of the academics in its three universities before 2003.

The biggest problems, however, are still faced by universities in the five eastern *Länder*. A huge injection of federal funds since German reunification in 1990 has improved the quality of computers and equipment in some scientific areas to Western standards. But in others, such as nanotechnologies, the east remains woefully behind, says Schipanski, herself former rector of the eastern University of Ilmenau. She describes the state of university buildings in the east as disastrous.

The *Länder* are also handling their universities' structural problems differently. The federal government would like all universities, which carry out one-fifth of all German research, to increase their commitments to research assessment. But its only influence is through the *Hochschulrahmengesetz*, a federal 'framework' law that gives guidance to the *Länder* for their own detailed university laws.

Jürgen Rüttgers, the research and education minister, has long tried to modify the framework law to reduce university bureaucracy and to require universities to assess their research regularly and distribute budgets according to the conclusions of such exercises. But his efforts have stumbled on the issue of student fees, as he refuses to

accept parliamentary demands for a clause forbidding their introduction by *Länder*.

Some *Länder*, however, have shown willingness to accept the spirit of the revisions to the law, even before they are approved by parliament. Lower Saxony has set up a scientific commission, headed by Wolfgang Frühwald, the former president of the Deutsche Forschungsgemeinschaft, to assess research performance in its 12 universities. And Bavaria intends to have research at its ten

Battle to boost the status of French universities

[PARIS] France's centralized universities are often characterized as having bursting classrooms, high teaching loads and a poor research base relatively isolated from the industrial world. This image is a far cry from the MIT-style research and technology university held up as an example by Claude Allègre, the minister for education, research and technology.

But Allègre's allusion to MIT (the Massachusetts Institute of Technology) reflects his conviction that universities should be the linchpin of the research and training base. Until now, French universities have played second fiddle in science to the public research organizations, and have been the poor relation of the elite *grandes écoles* of engineering.

Many argue that the predominance of the research agencies and *grandes écoles* has served France well, particularly at a time when the focus was on a few *grands programmes technologiques* in nuclear science, aeronautics and space. But there is a growing feeling that this system is less appropriate to today's challenges, such as biotechnology, robotics and computing, which require the bringing together of scientists, companies, well trained students and entrepreneurs.

Research funding in France is concentrated in the research organizations, such as the Centre National de la Recherche Scientifique (CNRS), the national biomedical research agency, INSERM, and the Atomic Energy Commission. The best funded university laboratories are mixed institutions that also house groups belonging to these research agencies.

Allègre wants to reduce this disparity by enforcing a system of four-year 'contracts' between his ministry and the universities that was created by Lionel Jospin during his spell as education minister from 1988 to 1992.

Many of the laboratories of CNRS and other research organizations are housed in universities. Until recently, control of these mixed laboratories was exercised almost entirely by their parent agencies. The new

universities evaluated by external experts, and to distribute research money on the basis of the results (see *Nature* 390, 210; 1997).

Daxner believes that the problems of universities are more likely to be resolved by demography than by legislation. Half of Germany's professors will retire before 2005, taking with them much of the conservative mentality that has resisted change. "What a unique chance for a new beginning," he says. Quirin Schiermeier & Alison Abbott



Allègre: looks to MIT as the way forward.

system of contracts gives the universities joint control, requiring agreement on research strategies and goals, as well as on the staff, resources and collaboration needed.

"The system of contracts is good," says Henri-Edouard Audier, a chemist at the

Ecole Polytechnique near Paris and a member of the board of a researchers' trade union, the SNCS. He says that contracts could create real synergies in research planning between the universities and research organizations. "There is a greater integration of universities [into the research landscape]," says Jacques Sevin, CNRS's director of strategy and programmes.

Another disparity between the universities and the research agencies is that agency scientists carry out research full-time and do not have to teach, whereas university staff often have almost no time for research. Despite much talk and exhortation, mobility between the universities and the agencies remains low. Fewer than 70 of the CNRS's 12,000 researchers moved to work within the universities last year, for example.

To reduce the differences, one idea being floated by the ministry is to end the separate recruitment of university and CNRS researchers. Researchers recruited to the agencies would be given the single joint status of lecturer/scientist. The time they devoted to research and teaching would be adjusted during their careers, with staff spending their early years working full-time in research, and taking on greater teaching loads later.

Few dispute that France has produced some excellent 'research universities' — for example, at Strasbourg, Compiègne and Grenoble. But the university system as a whole is in need of reform. Allègre, who had a distinguished research career as a university geophysicist, knows as well as anyone that there are no miracle remedies.

All students who pass the baccalaureate, the school-leaving examination, automatically qualify for higher education, and student numbers have leapt from 1 million in 1980 to more than 2 million last year. But four out of ten students drop out without a degree, while the growth in numbers has outpaced funding, leaving many universities overcrowded, short-staffed and in a dire state of decay (see *Nature* 385, 571; 1997).

Yet selection on merit has traditionally been fiercely resisted by students and faculty members alike as threatening the cherished principle of open access. This is ironic, given that France in fact operates ruthless selection of school-leavers, with a tiny élite being creamed off to the *grandes écoles*, leaving the majority to make do with a state degree from a university. But, while being excluded from a *grande école* is accepted, being excluded totally from higher education is not.

This quest for a form of *égalité* has also produced a centralized state system, with a 1984 law fixing national norms for everything from the make-up of university councils to course structures. This has resulted in a state degree from a leading research university such as Paris-Sud or Strasbourg having the same value as that of a lesser institution.

But attempts by successive conservative governments to give the universities a greater say in their affairs have come unstuck amid protests from students (see *Nature* 364, 5; 1993). The students argue that this policy would lead to institutions charging differential course fees, and would be a slippery slope towards an Anglo-Saxon system of competing universities with restricted entry.

Allègre has so far refrained from seeking a change in the law to break the state's grip on the universities (although he has transferred to them responsibility for recruitment from national committees). Instead, he has focused on strengthening the universities themselves. The government's creation of almost 3,000 university research posts this year will reduce teaching loads and boost research capacity. Demographic trends should also help — the student population is expected to plummet.

Allègre is fond of saying that French universities have "won the battle of quantity, now it is time to win the battle of quality". Indeed, Hélène Lamicq, the president of the University of Paris XII, points out that universities deserve credit for having successfully absorbed the massive increase in student numbers.

She is also convinced that, through negotiation, students could be persuaded to accept the increases in course fees that universities argue are needed if they are to improve facilities — provided the increases were across the board. Allègre intends to raise the issue of fees at a National Assembly debate this year. **Declan Butler & Eric Glover**

Britain faces research funding dilemma

[LONDON] At the beginning of a year that will bring some of the biggest changes in British higher education since the mid-1960s, universities are waiting nervously to see what the recently elected Labour government expects of them for the new millennium.

Despite the doomsayers, research in British universities is flourishing. A survey in 1997 by Sir Robert May, professor of zoology at the University of Oxford and the government's chief scientific adviser, showed that in many fields the quality of research is second only to the United States.

Some see this as a reflection of the success of mid-1980s reforms, in particular rigorous research assessment exercises used by the government as the basis of distributing money to institutions. But others warn that it also reflects relatively high past investment — investment now threatened by government parsimony.

Many such concerns were highlighted in what is inevitably seen as the most important policy document since the Robbins report of the 1960s — the report of the National Committee of Inquiry into Higher Education, under Sir Ron Dearing.

Most publicity surrounding the Dearing report concentrated on the way in which it paved the way for the introduction of student fees, ending the tradition of free university education still cherished by many other European countries. But the report contained other important recommendations for university research.

The government is due to respond with a policy statement in the next few months. An indication of the importance it attaches to university research will lie in its response to the report's complaint — almost universally endorsed — that the main problem facing researchers is lack of funding for infrastructure. Dearing recommended an extra £110 million (US\$184 million) a year, although the Committee of Vice-Chancellors and Principals (CVCP) calls this a "serious underestimate".

Sir Aaron Klug, president of the Royal Society, is one of many who have recently warned that unless adequate money for overheads is provided for in grants from research councils, cutbacks are likely in the overall volume of research that could have a crucial impact on the economy. "This small sum could make a huge difference to the health of the science base," he says.

Dearing did not suggest any basic change to Britain's cherished 'dual support system', which is designed to ensure a balance between 'core' funding provided through the four regional higher education funding councils and project funding through research councils.



Dearing: reaction to his report will produce far-reaching educational reforms.

But, in a further step away from the concept that all universities should automatically qualify for 'core' research support — and thus towards US-style research universities — it recommended raising the level of scientific recognition that individual departments must meet to qualify for such support.

Some are arguing that the committee may not have gone far enough. The Council of Science and Technology, the government's top independent advisory committee, says that "over time the qualifying standard may have to be raised further to ensure that the best groups are funded at an adequate level".

But others are more cautious. The CVCP — perhaps inevitably for a body that represents the collective interests of higher education institutions — agrees with Dearing's suggestions, but argues that the present level of selectivity in funding is broadly acceptable. "We do not want to see an over-concentration of research in too few universities," it says.

If the government response is expected to indicate how far it is prepared to go down the selectivity road, equally critical is how much it is prepared to nurture creativity among individuals and groups that have yet to establish a robust research track record.

Dearing suggested that university departments that agree not to bid for core research funding should be compensated with £500 for each faculty member to encourage independent 'scholarship'. This has been widely criticized for lack of focus. But a modified scheme, in which the money would be pooled and then distributed by competition, is said to be gathering support — and could play a key role in ensuring British research is continuously supplied with the new blood it needs. **David Dickson**

US universities are in flux, not crisis

[WASHINGTON] The past year closed to the sound of alarm bells alerting US policy-makers to the allegedly perilous state of higher education. But the research universities that occupy the pinnacle of the system say that they are strong enough — financially and intellectually — to weather the storm.

The Cassandras have certainly been out in force. Speaking last month at the National Science Foundation, Chang-Lin Tien, until recently chancellor of the University of California at Berkeley, argued that “time was running out” for the “very needy, very stressed” higher-education system to diversify its funding base, improve teaching and reform its tenure system.

And a report in the summer by the Council for Aid to Education, a subsidiary of the Rand Corporation, found that “the present course of higher education, in which costs and demand are rising much faster than funding, is unsustainable”.

The report, *Breaking the Social Contract: The Fiscal Crisis in Education*, called for “increased public funding of higher education and wide-ranging institutional reforms”, and for funds to be directed to the top-ranked research universities.

But a visitor to, for example, the vast oceanside campus of the University of California at San Diego, or the superb new Uni-



Boundless possibility: Research on the University of California's San Diego campus has blossomed.

versity of Arizona in downtown Tucson — two of the establishments that have evolved from near-obscurity to world-class status in the past 25 years — will not find an atmosphere of crisis. It is, instead, one of stunning, almost boundless possibility.

And among university administrators and senior government officials, the view that the system is in deep trouble remains a minority one. “I wouldn’t call it a crisis,” says Joe Bordogna, deputy director of the National Science Foundation. “There are changes going on and the universities are all having trouble adjusting, some more than others.”

Cornelius Pings, president of the Association of American Universities, which represents 60 leading research universities in both the public and private sectors, is similarly relaxed. “I certainly don’t sense any crisis, although I sense some increasing pressures and anxieties,” he says.

The financial pressure on universities involves distinct problems with each of the three main sources of funding — research funds from the federal government, general support from the 50 states, and earnings from teaching hospitals.

After four decades of substantial growth, for example, research funding from the Department of Defense (DOD) has stagnated since 1990, while support from the National Science Foundation has increased only very slowly.

But the budget of the National Institutes of Health, which pay for more university research than DOD and NSF combined, has continued to grow. And the threat of cuts of up to one-third in federal research support — very real two years ago — has also evaporated. “There were dire predictions that the research money would dry up,” says Pings. “It hasn’t.”

On declining support from state governments, the Council for Aid to Education report says that if trends continue, tuition costs would double by 2015, and the universities would face “a catastrophic shortfall in funding”.

But public and private universities are acutely conscious of the need to compete for middle-class students whose parents are too well-off to qualify for needs-based scholarships, but not rich enough to pay six-figure sums for their children’s education. Several, including Carnegie Mellon, New York University and the vast University of California system, have either frozen this year’s fees or raised them by less than in previous years.

Robert Shelton, for example, vice provost for research at the University of California, points out that it has been able to pin its undergraduate fee at \$4,800 a year for in-state students for several years.

The third and most vigorous pressure faces medical centres, where health-insurance companies are increasingly reluctant to pay for people to occupy research hospitals’ expensive beds. Mergers of important institutions, as well as closures of hospital beds,

Reforms get a rough ride in Italy

[MUNICH] Italian universities — like those in Germany and Japan — suffer from a lack of internal competition for research funds, and are further disadvantaged by the fact that they have a smaller pot of public research money.

At the same time, full professors remain powerful, and have for many years successfully opposed the introduction of a more accountable system of allocating research funds, including compulsory peer review.

Until last year, national research grants were allocated by discipline-based committees whose members were elected by the academic community. Most appeared to be concerned primarily with avoiding conflict by ensuring that no scientist was left empty-handed.

The government wants to exercise greater control over

the allocation of public research funds, to ensure both that more ‘transparent’ systems are introduced and that these funds are concentrated on fewer projects, in areas of strategic importance. It would like to see a general improvement in the quality of research in universities.

To achieve this, it has introduced a law dissolving the committees and replacing them with a single grant committee composed of five ‘wise men’ appointed by the research minister, Luigi Berlinguer. The five have now completed their first round of applications.

Despite concern in the academic community that five individuals would lack the breadth of expertise to select referees for the thousands of applications received, results suggest that concentration of research funds has for the first time

been achieved.

The new grant system places more responsibility on individual universities, which have been slow to embrace the autonomy granted to them in the late 1980s. Now, only projects selected by the individual universities for internal funding are eligible for a top-up from national research funds.

So far, however, the government has failed to increase its control over university grant money distributed by the CNR, Italy’s national research council. It wants eventually to remove CNR’s role as a grant agency, and transfer the moneys to the general universities fund, over which it has more direct control.

Berlinguer has proposed a change in CNR’s rules to allow this, but has so far been unable to overcome the inevitably strong resistance.

Alison Abbott

have swept across the system. But Bordogna points out that no medical centre in the United States has been closed: indeed some, such as the University of California at San Francisco, plan significant expansions.

None of the financial pressures on the system appears strong enough to shake institutions that do good research and attract good students. And although administrators criticize academics for neglecting teaching, many institutions are making progress integrating teaching and research.

"There's a very different atmosphere from 15 years ago," says Bordogna. Senior staff, he says, give much more consideration to what younger members of departments think. It is hard to prove, but many long-

term observers of the system agree. "We are light-years ahead of where we were," says Paul Christiano, provost of Carnegie Mellon University in Pittsburgh, Pennsylvania.

At the same time, competition between institutions for both research funds and students continues unabated. "The diversity of institutions in this country has been a blessing and a source of strength," says Christiano.

A remarkable aspect of this competition is the success of public institutions. An extensive study of doctoral programmes by the National Research Council found the institution with the most departments ranked number one in their disciplines is the publicly funded University of California at Berkeley.

The complaint persists that too many smaller universities are trying to do too much. Bordogna says the United States can support only "50, or maybe 30" top research universities, and wishes the other hundred or so aspiring research universities would concentrate on teaching.

But there is no possibility of Washington selecting that élite. Even Pings, who represents most of the leading universities, recoils at the idea of closing the door on other schools. "There should be an open, dynamic, competitive system. If an institution can muscle in, it should be allowed to do so," he says. So long as that opportunity remains, the research university system in the United States is likely to remain the envy of the world. **Colin Macilwain**

Japan faces problem of science recruits

[TOKYO] Despite claims by Japan's ministry of education that the country has now caught up with the West in both the output and the quality of its research, its universities still face chronic problems of limited flexibility and conservative attitudes. Critics say that recent policy moves may only aggravate the situation.

A 'white paper' produced by the ministry last month highlights the fact that Japan was second only to the United States in terms of its total output of scientific papers in the 1996 index of the Institute for Scientific Information (ISI). It also claims that Japan ranks fourth in the 'quality' of its research as measured by the percentage of total citations accounted for by Japanese papers. However, an analysis last year by ISI based on citation impact, or average citations per paper, which is the more generally recognized method of measuring quality, placed Japan seventeenth and below the world average (see *Nature* 389, 113; 1997).

But even the ministry admits that the universities, where most of Japan's public research is carried out, face severe problems. A fall in the student population means that there are too many universities competing for fewer students. Faculty members are ageing — the number of university researchers in their twenties fell from 11.6 per cent in 1977 to 4.5 per cent in 1995 — and the ministry expects a severe shortage of young researchers in the near future if these trends continue.

Responding to this situation, the government last year unveiled a plan to double the number of postdoctoral fellows to 10,000 by the end of the decade as part of a five-year plan to boost science and technology. The education minister is also considering a plan to prioritize education in graduate schools, and to increase the number of postgraduate students from 170,000 to 300,000 by 2010.

But some university faculty members are concerned that they will have to lower their



A task for the future: how to attract and encourage able students into research.

standards to fill graduate student and postdoctoral positions, and that four years of university education may be insufficient to produce graduates of international standard.

Another problem is the fact that faculty positions at national universities are subject to government regulations aimed at restricting the number of civil servants. Recent reforms led by Prime Minister Ryutaro Hashimoto are intended to cut the number of civil servants even further, and so the number of faculty positions at national universities is likely to be reduced still more, says Yoshiki Hotta, director of the National Institute of Genetics in Mishima, south of Tokyo. "The plan will be meaningless unless the government increases the number of research posts at universities," he says.

Leading policy-makers such as Akito Arima, president of the Institute of Physical and Chemical Research and a driving force behind government science policy, believe that a recent proposal by the education ministry and the Ministry of International Trade and Industry to strengthen the links between universities and industry (see *Nature* 390, 105; 1997) will create a large demand for university researchers.

But some in industry are sceptical. "A

strict selection mechanism will obviously be at work when we recruit new researchers," says Susumu Nishimura, director of research at Banyu Pharmaceuticals in Tsukuba science city.

"We would be looking very closely at whether they had received high-quality teaching and carried out interesting and creative research."

Nishimura points out that "many postgraduate students and postdocs at Japanese universities spend most of their time doing menial jobs under more senior researchers". A priority for Japanese universities must be to create an environment in which creative and original minds will thrive, he adds.

The white paper recognizes the need for a routine, nationwide external review system. Although Japanese universities began self-assessment of their research activities in 1992, less than half have so far brought in external reviewers, and only about a quarter of these have made the reviewers' reports public.

Critics also point out that there is an urgent need to reform university management so that the results of external reviews can be used to bring about improvements. Most external reviews have so far had comparatively little impact, and key recommendations are often not implemented, particularly if they involve changing fundamental management practices.

A recent attempt by politicians to convert universities into 'agencies' that would have greater autonomy but also greater accountability was quickly killed by the universities themselves, and by leading proponents of reform, as potentially restricting their level of public support (see *Nature* 389, 897; 1997).

But many leading Japanese scientists recognize that unless Japan implements routine external research evaluation and new mechanisms for managing universities, it will be hard for universities to produce creative and innovative research. **Asako Saegusa**

CEA researchers protest over threat to postgraduates

[PARIS] France's Atomic Energy Commission (CEA) is in danger of losing its postgraduate students. Although no decision has been made, there is concern about proposals emerging from a 'roundtable' on scientific employment set up by Claude Allègre, minister for research, last summer.

One proposal is that an agency should take broad responsibility for managing the system of financing postgraduates working for PhDs, and that these should be confined to university laboratories and related units. Such a move would be a blow for the CEA: the number of 'thésards' in its laboratories doubled over 10 years, reaching 1,103 in 1996.

Allègre has specified that any reform will be introduced in stages over three years. But in protest at measures described as a threat to "asphyxiate" the CEA laboratories, a petition among researchers attracted more than a thousand signatures in two days.

Dolly clone institute wins research funding reprieve

[LONDON] The UK government has reversed a decision to terminate funding for the research at the Roslin Institute in Scotland

that led to the successful cloning of the sheep Dolly (see *Nature* 385, 810–813; 1997). The Ministry of Agriculture, Fisheries and Food, which had earlier argued that the research project had been successfully completed and was now ready to be taken over by the private sector, has agreed to help fund further work on improving the nuclear transfer technology involved to March 1999.

Meanwhile, the Department of Trade and Industry (DTI), responsible for Roslin through the Biotechnology and Biological Sciences Research Council, has warned scientists against carrying out experiments that might lead to 'experimental' human beings. Replying to a report by the House of Commons Select Committee on Science and Technology, the DTI says that it is important to acknowledge limits to the attempts of researchers to push back scientific barriers.

Canada backs down on NATO science plan

[PARIS] Canada last week backed down from its decision to stop funding the North Atlantic Treaty Organization (NATO) science programme. The unilateral decision plunged the programme into crisis by blocking approval of next year's budget by the NATO council (see *Nature* 390, 207; 1997).

Following Canada's move, the council approved NATO's civil budget, which

includes around \$40 million for the science programme, and promised to review the concerns that led Canada to withdraw in the first place — that the programme should be more oriented towards strengthening cooperation with Eastern Europe and Russia.

Nobel winner shares Japan's science prize

[TOKYO] Leo Esaki, president of Tsukuba University and a Nobel prizewinner, is one of the three winners of the 1998 Japan Prize which rewards scientists for practical contributions to society. Esaki spent many years at the IBM Thomas J. Watson Research Centre in the United States, and won a Nobel prize in 1973 for developing the Esaki diode. He returned to Japan in the early 1990s. His award recognizes his work on superlattice crystals, used in semiconductor lasers and other electronic devices.

The other two Japan Prize winners are Josef Schell of the Max Planck Institute in Cologne, Germany, and Marc van Montagu of the Flanders Interuniversity Institute for Biotechnology in Ghent, Belgium. They share the ¥50 million (US\$390,000) prize for Biotechnology in Agricultural Sciences for developing methods of inserting genes into plants to enable them to resist insects or disease.

US mentor scheme goes nationwide on the Net

[BOSTON] A US mentoring programme for women in science and engineering will go nationwide after a two-year pilot at Dartmouth College in New Hampshire (see *Nature* **382**, 383; 1996). MentorNet, based at San Jose State University in California, will pair women undergraduates and graduates with mentors in industry via the Internet.

Carol Muller, executive director of MentorNet, says the goal is to establish 250 pairs this academic year, with a big increase in later years. AT&T and Intel provided start-up funds, and further corporate partners are being sought. MentorNet is sponsored by the Women in Engineering Programs & Advocates Network, a nonprofit educational organization founded in 1990 to improve "academic and social climates" for women in science and engineering.

UK alters badger policy in search for TB data

[LONDON] The British government has agreed to stop the general culling of badgers, which are believed to cause bovine tuberculosis (TB). But badgers in bovine TB hotspots will continue to be culled under a scheme designed to provide controlled evidence of links with the disease. The move comes after a



report by a scientific review group headed by John Krebs, chief executive of the Natural Environment Research Council.

The report says that, although most available evidence suggests the badger (above) is a "significant source" of TB in cattle, its role cannot be quantified because of lack of data. But it says the disease can best be controlled by developing a cattle vaccine.

Staff objections fail to stop ANU changes

[CANBERRA] A restructuring of the senior management of the Australian National University in Canberra is to proceed despite objections from the boards of the Institute of Advanced Studies and the faculties (see *Nature* **390**, 435; 1997). The two boards representing academic staff had described the plans as "confusing". They want control of resources to be returned to the senior posts that administer the institute and the

faculties, with each post to be the same level as the new post of deputy vice-chancellor.

Vice-chancellor Deane Terrell has been given the support of the university council to reject the academics' arguments, and the vacancy for a single deputy vice-chancellor will be now advertised. But a heated debate between staff and management appears set to continue after Derek Robinson, a new academic member of the council, moved successfully for the issues to be aired at the full council meeting in March.

EMBL set to award its own doctorates

[MUNICH] The European Molecular Biology Laboratory (EMBL) in Heidelberg will in future be able to grant its own PhDs. EMBL's governing council approved the idea last month, and the laboratory will now negotiate formal recognition of the qualification with accreditation committees in member states.

Until now, students enrolled in EMBL's PhD programme, which has run since 1983, have had to register with a European university which confers the actual degree. EMBL has so far submitted more than 150 PhD theses to 76 European universities — with a 100 per cent success rate. In future, students will be able to opt for a degree from EMBL, from a university of their choice, or a joint university-EMBL degree.

Spotlight needed on Italian policy

Sir — The battle about the long-awaited reform of Italian research policy¹ is taking place behind closed doors, and the scientific community is in the dark. There is no indication that there will be a transparent system of peer review² in the near future. Indeed, any discussion of reform will continue to be abstract and ineffective so long as there are no objective indicators of scientific performance.

In the past 15 years, panels and committees have been set up to evaluate the scientific performance of government-funded research organizations, but neither detailed reports nor action have followed. Even the recent working group of the Ministry of Universities and Research³ has provided no more than a laudable declaration of intent and a comprehensive catalogue of scientific centres, research activities and numbers of employees.

The 71 pages of graphs and comments provide no data on cash flow compared to scientific production per research centre and scientific project, or basic information on how, where and with what results money is spent.

This is unfortunate; a fast, reasonably objective and cost-effective way of measuring the scientific performance of research institutions is to use bibliometric indicators⁴. Statistics such as the number of publications and their relative impact factors can show quite surprising results. For instance, in a pilot study⁵ we have scrutinized the publication output of four main categories of scientific institutions in the field of environmental research: universities, CNR (National Research Council), ENEA (National Agency for New Technologies, Energy and the Environment), and all the other smaller organizations.

From the results reported in Table 1, Alison Abbott's claim that ENEA should "have less to fear" from scientific reform is disputable¹. In fact, we are still waiting to see the "thousand flowers" bloom as was claimed by ENEA's president, Nicola Cabibbo — the quality and productivity of ENEA research in the environmental field are among the lowest to be found in Italian science.

The Ministry of Universities and Research could easily extend this study to other scientific fields and compare the research performances of different scientific institutions and research fields both within Italy and perhaps with those of other European countries. We are aware that bibliometric indicators are not exempt from criticism^{6,7} and that they should be used with care in conjunction with

Table 1 Performance of Italian scientific institutions in the environmental research field

Institutions	Article share	Average No. of citations per article	Average impact factor
Universities	55.3%	3.94	1.25
CNR	15.9%	4.68	1.24
ENEA	1.9%	2.22	1.01
Other	26.9%	3.89	1.34

These statistics⁸ have been computed from 4,858 articles published between 1981 and 1995 in 231 leading journals selected from environment-related categories of Scisearch⁸.

strategic, cultural and political considerations. But it cannot be denied that they can bring objective and valuable information to bear on the waste of resources within the system.

Yet, after years of discussions, the bogeyman of peer review based on bibliometric indicators still comes up against ingrained fears and strenuous resistance, because of its potential to reveal structural inefficiencies and, consequently, to trigger disturbing conflicts and bring the discussion under the spotlight of public attention.

Giulio De Leo
Gianmarco Paris
Marino Gatto

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: paris@dsa.unipr.it

Paolo Menozzi
Dipartimento di Scienze Ambientali, Università degli Studi, Viale delle Scienze, 43100 Parma, Italy

Life, love and soil

Sir — Daniel Markewitz¹ has objected to the use of the term 'soil' for material recently studied remotely by Mars Sojourner, because he thinks soil should be defined as a medium for life. But there is a very compelling reason why the US National Aeronautics and Space Administration and many scientists choose to talk of soil on the surface of the Moon and Mars. They wish to be understood. Soil, like love and home, is difficult to define, but is, at the same time, a familiar and comforting concept².

A geological as opposed to an engineering or agricultural definition of soil is material altered in place at the surface of a planetary body by physical, chemical or biological means³. This

Sir — Biologists in the university sector share Ben Mifflin's pride and pleasure that fellow biologists in 15 UK research institutes have average citation rates for their papers that exceed those of their peers in the United States, Canada, Germany and France⁴.

That, however, is the correct comparison, institute with institute, like with like. Mifflin's attempt to show that biologists in UK institutes are somehow more successful than biologists in UK universities is misguided (and invidious). He does not compare like with like.

The two types of organization and their staffs have different roles in our society and contribute differently to the progress of biological science.

Peter Ayres
(Chair, Heads of University Biological Sciences)
Department of Biological Sciences, Institute of Environmental and Nature Sciences, Lancaster University, Lancaster LA1 4YQ, UK
e-mail: p.ayres@lancaster.ac.uk

- Abbott, A. *Nature* **388**, 609–610 (1997).
- Abbott, A. *Nature* **383**, 567 (1996).
- Berlinguer, L. Linee per il riordino del sistema nazionale della ricerca scientifica e tecnologica. Università e Ricerca **3** (1997).
- Mifflin, B. *Nature* **390**, 12 (1997).
- Gatto, M., De Leo, G., Menozzi, P. & Paris, G. Fifteen years of environmental research in Italy: an analysis of international publications produced between 1981–1995. Proc. S. It. Ecol. 8th Congress (1997).
- Motta, G. *Nature* **376**, 720 (1995).
- Metcalfe, N. B. *Nature* **376**, 720 (1995).
- Institute for Scientific Information (ISI) Topical Citation Report 1981–95 (ISI, Philadelphia, PA, 1996).

definition of soil introduces the idea of alteration in place as a distinction between soil and sediment. This important distinction is deliberately blurred by terms such as regolith and gradation, also used in planetary geology⁴, especially where only remote-sensing data are available. But in martian images, sediment, which is bedded, can be distinguished from soil, which is clayey and salty³.

Gregory J. Retallack
Department of Geological Sciences, University of Oregon, Eugene, Oregon 97402-1272, USA
e-mail: greg_retallack@cmail.uoregon.edu

- Markewitz, D. *Nature* **389**, 435 (1997).
- Hole, F. *Geoderma* **25**, 75–112 (1981).
- Retallack, G. J. *Soils of the Past* (Unwin-Hyman, London, 1990).
- Greeley, R. *Planetary Landscapes* (Unwin-Hyman, London, 1985).

Eighteen-ninety-eight and all that

Anniversaries of note this year range from the discovery of alpha particles and inert gases to a plague of medical breakthroughs, the creation of a health service to make them available... and the invention of an ice-cream freezer.

J. L. Heilbron and W. F. Bynum

*Before thy feet I fall
Lord, who made high my fate
For in the mighty small
Thou showd'st the mighty great.*

The mighty small was the mosquito's parasite, the mighty great the mechanism of malaria, the prostrate poet with the high fate Ronald Ross, recipient of the Nobel prize for physiology or medicine in 1902 and of an honour then even rarer for a man of science, a knighthood, in 1911.

Ross wrote the verse on 21 August 1897, the day after he had observed in the stomach wall of a mosquito of a new type for him, *Anopheles*, some pigmented granules similar to those found in the blood of malaria victims. The Indian Medical Service for which he worked was less than ecstatic about his discovery, and refused him the time and resources to follow it up; and it was not until 1898 that he managed to work out the life cycle of the parasite responsible for the transmission of malaria in birds. In that year, Italian parasitologists, led by Giovanni Battista Grassi, demonstrated that the same cycle holds for human beings.

We choose as the anniversary of the year the discovery 100 years ago of the intricate relationship between man, malaria, mosquito and plasmodium. While demonstrating the intricacies of the anniversarial art (Ross always preferred 1897), our choice offers the opportunity of healing retrospectively the bitter priority dispute between these two headstrong men.

There is an alternative anniversary, which many might prefer. During 1898 a woman working under conditions almost as unfavourable as Ross's took to pieces a ton or so of uranium ore in pursuit of what she and her husband called "radioactivity". By the middle of the year Marie Curie had isolated a new element she named "polonium" after her native country. Around Christmas she and Pierre Curie announced the existence of a radioactive element more powerful than uranium or polonium, termed "radium", which seemed to them to violate the most sacred principles of physics, even the "theorem of Carnot", better known as the second law of thermodynamics. Like Ross's discovery, that of the Curies had medical consequences; but, whereas the extermination of *Anopheles* saved human life, the application of radium directly to human bodies probably did more harm than good.

As secretaries of nature (to use the noble



Implicated insect: *Anopheles* malarial mosquito.

term Henry More bestowed on René Descartes), as well as theorists of commemorations, we find the case for the choice of malaria irresistible. Epidemic diseases run through this year's crop of anniversaries. The old chronologists who saw dangers in years divisible by 100 were surely right to suppose an intimate connection between the course of human affairs and the number of human digits. As to the mechanism of the connection, we confess that, as yet, we know no more than they did.

We begin as usual with events eligible for centennial and sesquicentennial celebrations, back to 1098, after which we return to our century to mark notable events in science 50 and 50 ± 25 years ago. We use our usual anniversarial abbreviations: ¢ represents 100 years, and \$ indicates a Nobel prize with specification of year. Thus \$1902 signifies Ross's Nobel prize for physiology or medicine, and \$1903 the Curies' prize for physics.

1898: Pathogens and particles

Ronald Ross's father made him a doctor. He would have preferred to be an artist or literary man, and he spent his leisure during his first spell as a military medical officer in India teaching himself mathematics. During a furlough in England in 1894, he met Patrick Manson, whose *Tropical Diseases* of 1898 is also worthy of centennial notice. Manson demonstrated to Ross the causative organism of malaria, discovered in 1880 by the French military doctor Charles Laveran, and shared with Ross his (Manson's) conviction that the mosquito was somehow implicated in the spread of the disease. Ross returned to

India determined to prove the hypothesis correct. The search lasted four years; henceforth, Ross devoted his energies to the extermination of the mosquito.

Another insect with a similar malevolent life style was inculcated in 1898, also on evidence collected in India, when Paul-Louis Simond conjectured, correctly, that rat fleas transmit bubonic plague to humans. While on the subjects of blood and incrimination, we should not withhold the information that in this same year, 1898, a jury in the United States made the first conviction based on the identification of blood stains by a chemical test.

Radium and radioactivity had their congeners too. It was in 1898 that Ernest Rutherford (\$1908, 0.9¢), analysing the radiation from uranium and thorium, discovered that there were at least two types of rays, one heavily ionizing and little penetrating, which he called "alpha", the other, of longer range and lesser effect, termed "beta". His gradual unmasking of the character of alpha particles led him a decade later to the discovery of the nuclear atom. He had drifted into the study of waves and ions at the Cavendish Laboratory in Cambridge, England, at which he had arrived in 1895 intending to improve the detection of radio waves. He would have had serious competition in Guglielmo Marconi (\$1909), who in 1898 managed to transmit a radio signal across the English Channel.

One hundred years ago Martinus Willem Beijerinck, a botanist and chemical engineer, identified the pathogen of the mosaic disease in tobacco plants. It was a new sort of agent, which kept its activity after diffusion through a filter that stopped all known microorganisms. While Beijerinck was studying this surprisingly vigorous "virus", William Ramsay (\$1904) and Morris Travers discovered the surprisingly inert gases neon, krypton and xenon, and James Dewar liquefied hydrogen, at temperatures at which almost nothing ever happens.

Several Nobel prize-winners would have celebrated their one-hundredth birthdays this year. They include Howard Walter Florey, 'ennobelled' for the discovery of penicillin and its curative effects (\$1945); Karl Ziegler, for his work on polymers (\$1963); and Isidor Isaac Rabi, for his method of investigating the magnetic properties of nuclei (\$1944). Rabi was born on 28 July, as was Charles Townes (\$1964), albeit in a year, 1915, not entitled to anniversarial notice according to our usual rules. But in our constant effort to enhance our art, we encourage the celebration of all Nobel prize-

winners born on the same day if any one of them qualifies for a centennial. We have not gone so far as to admit as winners those who would have obtained a prize if only they had been right, such as Trofim Denisovich Lysenko, who, however, qualifies by birth for a centennial party, should anyone care to give him one.

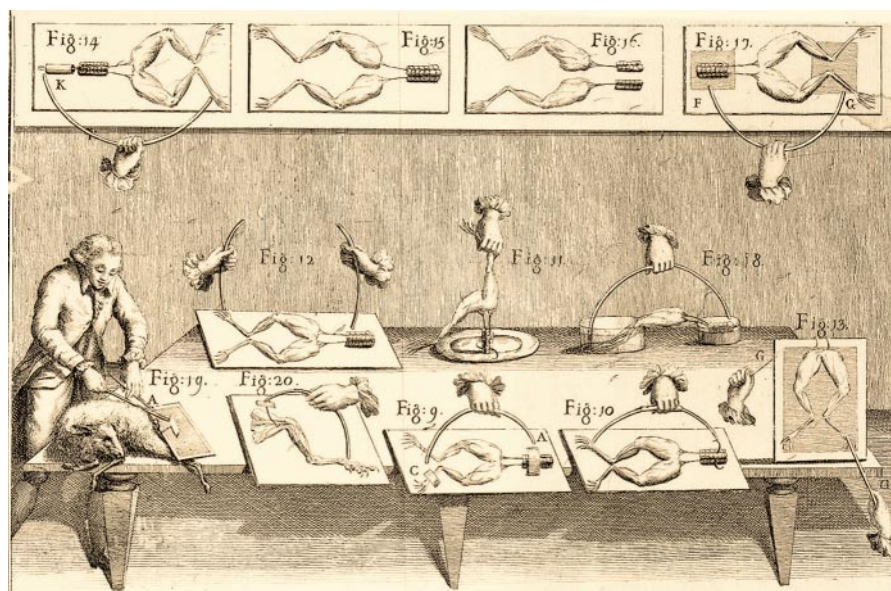
Other tempting prospects are two of Albert Einstein's co-workers, Leopold Infeld (died 1968, 0.3¢) and Leo Szilard. Infeld, a Polish mathematician, collaborated with Einstein on the relativistic motion of heavy bodies and on the pot-boiler (literally that, to keep Infeld going) *Evolution of Physics* (1938, 0.6¢). Szilard, a Hungarian physicist, collaborated with Einstein on an electromagnetic pump for refrigerants that is now used in nuclear reactors, and also on the famous letter alerting President Franklin D. Roosevelt to the possibilities of nuclear explosives. When the United States succeeded in making atomic bombs, Szilard campaigned in vain against their use.

Among those who have been dead for a hundred (and only a hundred) years are Henry Bessemer, who invented his converter to make steel strong enough for rifled artillery pieces; Sir William Jenner, whose differentiation of typhus and typhoid fevers unfortunately did not provide the requisite therapeutic breakthrough when Prince Albert was fatally stricken with typhoid fever in 1861; Johann Jakob Balmer, a teacher in a girls' secondary school in Basel, who made his reputation for a bit of numerology by connecting the frequencies of spectral lines in hydrogen; and Charles Lutwidge Dodgson, a mathematician who made his reputation as Lewis Carroll by writing stories for little girls.

1848: Cold comfort

In this year of still-born democratic revolutions, cholera epidemics, and the *Communist Manifesto* of Karl Marx and Friedrich Engels, Louis Pasteur imperturbably separated tiny crystals of racemic acid by hand into two types, one capable of rotating the plane of polarized light to the right and the other to the left. The discovery of these isomers explained why tartaric acid, chemically identical to racemic acid, was optically active but racemic acid was not.

Thinking bigger but less successfully, Julius Robert von Mayer (died 1878, 1.2¢), one of the principal discoverers of the conservation of energy, traced the source of the Sun's heat to the continuous impact of meteors on its surface. This suggestion was rejected by William Thomson, later Lord Kelvin, another and greater contributor to the foundations of thermodynamics, on the grounds that the resultant accretion of matter would have produced an unobserved change in the motions of the planets. Thomson is eligible for a sesquicentennial observance this year



Electric performer: Galvani experiments on frogs (1792).

for his enunciation, in 1848, of the principles of the absolute scale of temperature, called after his lordly name.

We are glad to record that while the Scot Thomson theorized about temperature, an American did something about it: 150 years ago William G. Young, of Baltimore, Maryland, patented the ice-cream freezer.

A deeper freeze chilled German-born Caroline Herschel, who collaborated in the all-night vigils of her brother William, the discoverer of Uranus, and received the coveted gold medal of the Royal Astronomical Society for her reduction of his observations on nebulae and star clusters; Jöns Jacob Berzelius, Swedish one-time dean of European chemistry; the British railroad engineer George Stephenson; James Cowles Prichard, English physician, psychiatrist and ethnologist, who died from an infection caught while inspecting lunatic asylums; and, by his own hand, Horace Wells, a great benefactor of the human race, not through his suicide but by his invention of surgical anaesthesia.

Those born in this revolutionary year included Roland Eötvös, the Hungarian physicist whose precise measurements of gravitational mass played a part in the gestation of the general theory of relativity; Albert Grimaldi, Prince of Monaco, a great promoter of oceanography, whose discoveries included new species of giant squid found in the stomach of a captured whale (an expensive way to fish, even for a prince); Henry Rowland, inventor of the optical grating with which spectroscopists prepared their contribution to the theory of atomic structure, and also a ferocious fighter for the support of basic science; the American Association for the Advancement of Science, a generation after its British and German counterparts; the Société de Biologie, whose foundation its vice-president Claude Bernard (died 1878, 1.2¢) celebrated by discovering the glyco-

genic function of the liver; the General Board of Health in England, a belated response to cholera and other epidemic diseases; and the University of Wisconsin.

To be on the safe side, we mention that in 1848 Linus Yale (died 1868, 1.3¢) patented his rotating-cylinder lock.

1798: Forces, friction and frogs

It is possibly no more than a coincidence that fifty years before the birth of the precise metrologist Roland Eötvös, the Honourable Henry Cavendish made the first measurements of the sort that were to establish Eötvös's reputation. Cavendish succeeded in detecting directly the force of gravity between two stationary and two fixed lead masses, and in deducing the density of the Earth as 5.5 times that of water. An amalgamation of the commemorations of Cavendish and Eötvös is in order.

The year 1798 was one for ponderous science. While Cavendish was weighing the Earth, Benjamin Thompson (Count Rumford) published his *Enquiry Concerning the Source of Heat which is Excited by Friction*. As every schoolboy knows, Thompson reasoned from his observations of the heat generated in the boring of muzzles of large artillery pieces. His conclusion, that heat is a form of motion, did not carry conviction in his day. It may be worth the attention of philosophers that no contemporaries of Cavendish or Thompson cared to repeat their experiments.

The Norwegian surveyor Caspar Wessel (died 1818, 1.8¢) invented the method of representing complex numbers in the plane usually credited to Karl Friedrich Gauss or Jean Argand. This may not be an instance of the Matthew effect (to all those who have, more will be given) but a consequence of his writing in Danish. His complex invention almost deserves a double anniversary, as it

was discovered three years short of a century after it was published.

Death took the destroyer of frogs, Luigi Galvani, a physician of Bologna. An elegant performer at the carnival anatomies executed before the élite of his city, he discovered the extraordinary sensitivity of disembodied frogs to electricity. His resourcefulness in defending his theory that frog electricity differed from the ordinary kind stimulated Alessandro Volta to refine the counter position. Thus Galvani must be credited with a good share of the glory that, in the usual way, has accrued to his victorious opponent.

Two hundred years ago Thomas Malthus published his *Essay on the Principle of Population*, which proved that starvation was inevitable on the fringes of society because the growth of the food supply can never keep up with the increase in population, as one enlarges arithmetically, the other geometrically.

On a more optimistic note, Edward Jenner (died 1823, 1.75¢, no relation of William Jenner; see 1898) published, at his own expense, *An Inquiry into the Causes and Effects of the Variolae Vaccinae*, and so made public his experiments with cowpox begun two years previously. Jenner's original communication had been turned down by the editors of *Philosophical Transactions*, although his earlier observations on the breeding habits of the cuckoo had already secured his election to the fellowship of the Royal Society. Further to good health, we should note that the US Marine Hospital Service, forerunner of the US Public Health Service, has claims this year to a bicentennial.

1748: Earthy matter

This was the year when James Bradley, Astronomer Royal of England, and ranked by those who knew him with Ptolemy and Johannes Kepler, made public his discovery of the nutation of the Earth. This bobbing of the Earth's axis, a consequence of the theory of universal gravitation, arises from the pull of the Sun and the Moon on the Earth's equatorial flab. Together with the discovery, also by Bradley, of the aberration of starlight and with the growing conviction that the Earth's axis also undergoes a secular change in orientation, the nutation created great difficulties for the defenders of geostatic astronomies. No doubt Bradley's work helped Pope Benedict XIV to strike the blanket proscription against Copernican books from the Roman Index in 1758 (2.4¢).

Two books of 1748 gave the Pope other things to fret about. Benoît de Maillet's (died 1738, 2.6¢) posthumously published *Telliamed* (the curious title is the author's name spelled backwards) attributed the present geological features of the Earth to the slow retreat, over the past two billion years, of a once universal ocean, posited life to the generation of tiny seeds from outer space, and

argued that humans had developed from mermen and mermaids who once inhabited the seas. Meanwhile Julien Offray de La Mettrie's *L'Homme Machine* marshalled the evidence favouring the proposition that human beings are merely complicated machines, and that consciousness is simply a physiological function of the brain. The Pope undoubtedly slept easier when, three years later, the machine that was La Mettrie ceased to operate while processing a surfeit of pâté.

Subsequent study of *L'Homme Machine* was abetted by the invention of the eudiometer, which indicated the suitability of air for respiration, by Stephen Hales, a village clergyman, who also measured the pressure of tree sap and founded pneumatic chemistry; and by the discovery of the diffusion of liquids across thin membranes, the exemplar of osmotic action, by the abbé Jean Antoine Nollet, a metropolitan savant and pedagogue, whose *Leçons de Physique* reached its sixth and final volume in the same year.

1698: Fiery matter

In this year Thomas Savery, a Devonshire man, patented a steam pump "for raising water or occasioning motion to any sort of millworks by the impellent force of fire". As straight a line as the history of technology afforded led from Savery's pump to the engine with separate condenser that James Watt patented in 1769 (2.3¢ to 0.4 per cent).

1598: Un-English ills

Unlike the sober Scot Patrick Manson (see 1898), George Whetstone (1544?–1587?) lived his life in the fast lane. He squandered his patrimony on riotous living, wrote bad verses and worse plays, sailed with Humphrey Gilbert in an abortive voyage to Newfoundland, and fought against the Spanish in the Low Countries. He also probably penned the first treatise in English on

diseases of foreign climates, *The Cures of the Diseased, in Remote Regions*, by G.W., published a few years after Whetstone's death, and exactly three centuries before Manson's own effort.

1498: Neither here nor there

There was the usual running to and fro. Christopher Columbus, still struggling to reach the East Indies by sailing west, stumbled across South America; Vasco da Gama got to the Indies, or anyway to India and Malabar, by going in the opposite direction; and John Cabot enjoyed a voyage along the coast of Florida.

1348: Plagues and poets

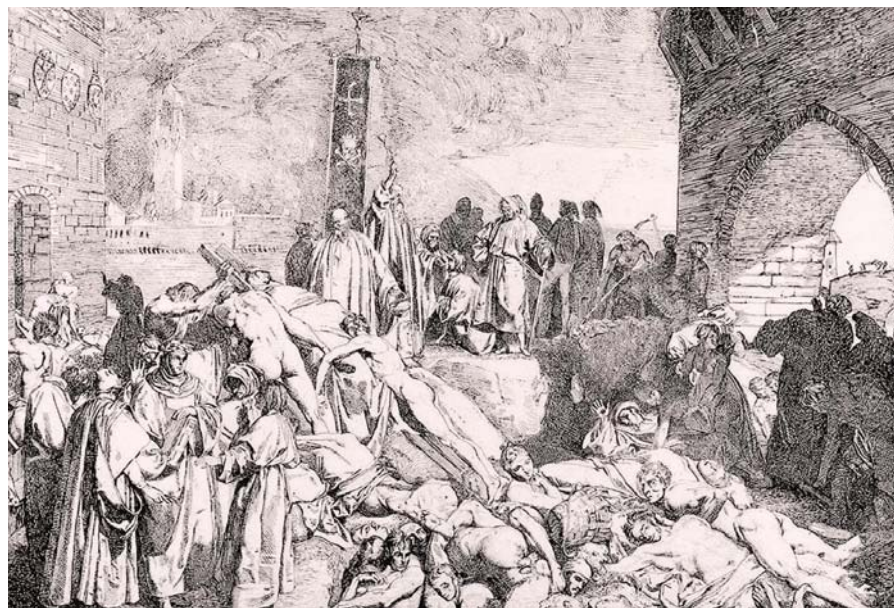
Malaria may have stimulated Ross to write a minor poem, but the Black Death provided the backdrop for an early Italian Renaissance masterpiece, Boccaccio's *Decameron*, begun in 1348 in a Florence in the grip of the most devastating epidemic of its history. Boccaccio believed that 100,000 Florentines died of the plague between March and June 1348, and between a quarter and a half of the population of Western Europe had perished from the disease by 1350.

1098: Nunsuch

Hildegard of Bingen was born in 1098, went into a nunnery at the age of eight, and died in a convent of her own founding 73 years later. In between, she wrote down what she saw and heard, concentrating on the visions she had had from her childhood. Although her inspired Latin had to be corrected by a collaborator, her cosmological and medical writings were so full of insight into the nature of things that even popes took notice.

1948: Health for all

Inspired should be a word used to describe the beginning of one of the greatest social



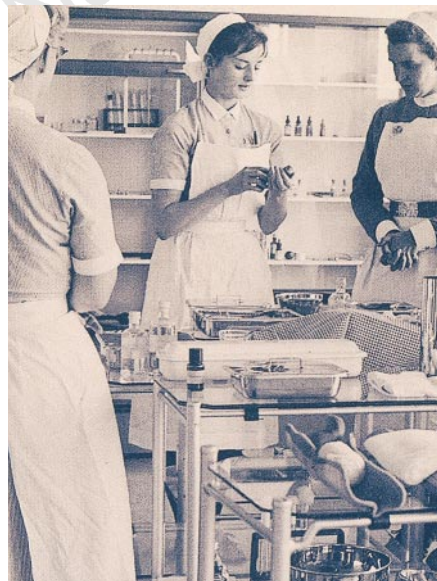
Fortune teller: scene from Boccaccio's *Decameron* of the plague of Florence (1348).

experiments of our century, Britain's National Health Service, which left the maternity ward on 5 July 1948. A look at health expenditure in any Western country after the Second World War should convince any economist or politician that the NHS has been, and is still, an excellent investment, deserving better than it generally gets.

No less than universal health was the object of the World Health Organization, established half a century ago and superseding the functions of the older League of Nations Health Organization and the Office d'Hygiène Publique. The eradication of smallpox in 1975 represents its greatest single achievement. Its assault on malaria, more contentious and much less successful, was stimulated by the availability of the pesticide DDT, developed by Paul Hermann Müller (\$1948). And that does not exhaust this year's malarial celebranda: in 1948 the anti-malarial drug Pentaquine went into clinical trial, and H. E. Shortt and Percy Garnham demonstrated the pre-erythrocytic phase of the disease in the livers of monkeys. It was obviously healthy work, as between them Shortt (1887–1987) and Garnham (1901–1994) lived for 193 years.

Among drugs that have stayed the course, but should not be celebrated with champagne, we mention Antabuse (tetraethylthiuramdisulphide), introduced in 1948 for the treatment of alcoholism. Its remarkable properties were discovered by the Danish pharmacologists Jens Hald and Erik Jacobsen, who believed it might discourage intestinal worms. While taking the drug to see if it was safe, they discovered after a few days that the beer they had with their lunchtime sandwiches made them violently ill.

An object that grew even faster than the NHS was the Universe. Its expansion was credited in 1948 to an initial Big Bang,



Born free: the NHS emerges as one of the century's greatest social experiments.

according to George Gamow, Ralph Alpher and Robert Herman, and to the continuous creation of space and matter, according to the optimistic account by Hermann Bondi, Thomas Gold and Fred Hoyle. The great 200-inch reflecting telescope on Mount Palomar in southern California came on-line in 1948 to gather a little more light on the subject from receding stars.

Berkeley's 184-inch synchro-cyclotron, fully demobilized after distinguished service in the making of atomic bombs, produced pions recorded by Eugene Gardner, who died shortly afterwards of beryllium poisoning contracted during the Second World War, and the Brazilian physicist Césaire Lattes, who had learned in England how to detect mesons. Like the 200-inch reflector, the cyclotron was paid for mainly by the Rockefeller Foundation. The odd size of the accelerator's pole pieces, 184 inches, was the closest the Berkeley cyclotroneers could come with commercially available steel to matching the diameter of the Palomar mirror.

The year was filled with the seeds of inventions without which we now cannot live. William Shockley (\$1956), Walter Brattain (\$1956), and John Bardeen (born 1908, 0.9¢, \$1956 and \$1972) made the first transistor, which revolutionized consumer as well as industrial and military electronics. Peter Carl Goldmark developed the first long-playing record, which, with transistor technology, made possible the blessing of unending piped music. Edwin Land invented the camera with the developer inside, and the Swiss engineer George deMestral thought up the bond that binds much of our centrifugal life together, Velcro.

It is a long way from the crude few silent frames of the first moving picture to the Land camera and wide-screen Technicolor films with stereophonic sound; and also from the short, choppy, dangerous first flight at Kitty Hawk to transcontinental jaunts on air liners. Yet Louis Lumière and Orville Wright, who, with their respective brothers Auguste and Wilbur, invented the cinematograph and the aeroplane, had seen all that by the time they died in 1948.

1948 ± 25: Compared and contrasted

Remaining in the comparative mode, we note that Wilhelm Conrad Roentgen (\$1901), the discoverer of X-rays, died in 1923. Fifty years later, medicine received the new diagnostic tools of CAT scans and NMR probes.

The paradoxes of microphysics dominated scientific thought in 1923. This was the year that Arthur Holly Compton (\$1927) publicized his discovery that hard X-rays, previously thought to be light-like, behave as particles in collisions with electrons; Louis de Broglie (\$1929) began to develop his idea that electrons can behave as waves; and Robert A. Millikan was rewarded (\$1923) for



Higher research: Skylab 2 space station.

his measurements of electronic charge based on the assumption that its carrier was particulate. By then physicists had grown accustomed to the paradoxes of relativity, among them the theory of the Michelson–Morley experiment, whose junior author, Edward W. Morley, died in 1923.

In 1973, space exploration dominated the popular imagination, and molecular biology was beginning to capture some of the allure of physics. But the physicists continued on their way, with the scarcely communicable discoveries that neutral currents in neutrino reactions required by electroweak theory exist, that quarks attract one another more powerfully when they are further apart, and that the Universe (see 1948) could have arisen *ex nihilo* by fluctuation of the quantum vacuum. Meanwhile, the US space agency NASA made history by putting into orbit three sky labs, in each of which three men performed physical and biological experiments in space. Astronomers discovered organic molecules in the head of a comet, suggesting that life might be found in improbable places, and Stanley H. Cohen (\$1986) and Herbert W. Bayer showed how to cut and paste bits of DNA and to reproduce the home-made strands, opening up the field of genetic engineering.

We note, as further evidence of the scientist's influence on all aspects of life, the passing of Selman Abraham Waksman (\$1952), the Russian–American microbiologist whose discovery of streptomycin prevented many a death from tuberculosis and other scourges, and of Robert Watson-Watt, the Scottish physicist whose work on early radar contributed to victory in the Battle of Britain. But life goes on: 1973 also saw computer-controlled labels and readers in supermarkets and tab openers on beer cans. □

J. L. Heilbron is at Worcester College, Oxford OX1 2HB, UK. W. F. Bynum is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BE, UK.

How black holes stay black

Reinhard Genzel

There is probably a massive black hole at the centre of our Galaxy, and perhaps at the centre of most galaxies. Their dimness surprises astrophysicists, but a possible explanation has been found in the behaviour of the plasma they consume.

In the theory of General Relativity, black holes are concentrations of mass so great that even light cannot escape from them. Paradoxically, in extragalactic astrophysics black holes are considered to be the best way to explain the large luminosities of distant active galactic nuclei, such as quasars. More recently, evidence has been accumulating that large black holes also exist in the centres of many, perhaps most, normal galaxies — but that they cause far less radiation than expected. In a paper to be published in the *Astrophysical Journal* on 10 January¹, Ramesh Narayan and co-workers offer an explanation of why one such candidate black hole, suspected to lie at the centre of our own Galaxy, is so remarkably quiet.

The paradox mentioned above is resolved when one realizes that black holes are completely 'black' only within the event horizon, the so-called Schwarzschild radius, which is three million kilometres for a million-solar-mass black hole. Outside this radius, light can escape. Massive black holes in galactic nuclei suck in nearby stars and interstellar gas through their enormous gravitational pull. On its way into the hole, the accreting matter (probably in the form of a rotating disk) heats up through viscous dissipation and so converts gravitational energy into radiation. Calculations show that in such accretion disks about 10% of the initial rest-mass energy is radiated away. This is the most efficient radiation source one can think of, nuclear fusion in stars being a distant second. For this and several other reasons the 'massive black hole accretion' model has been the standard explanation for the spectacular energy output of quasars.

During the past decade, astronomers have made observations at ever higher resolution to test this black-hole model in nearby, low-luminosity galaxies — the mass distribution is inferred from velocity measurements of gas and stars. Compact, dark mass concentrations between a million and several billion solar masses exist in about a dozen nearby galaxies². In the galaxy NGC4258 (ref. 3) and our own Galactic Centre⁴, we can now say with some confidence that the central dark mass cannot be anything but a black hole. In the case of the Galaxy, the evidence

for a 2.6×10^6 -solar-mass black hole is based on measurements of radial velocities and proper motions of stars as close as five light days from the suspected dynamical centre, which coincides with a bright, compact radio source, SgrA* (Fig. 1).

When one compares the nuclear luminosities to the accreting black hole model for about ten of the best black hole candidates, a surprising fact emerges: few of them radiate at maximum, quasar-like efficiency. Most nearby black hole candidates are remarkably radiation-deficient (L. Ho, personal com-

munication). The most extreme case is the Galactic Centre; it radiates only 10^{-8} to 10^{-7} of the maximum, so-called 'Eddington' luminosity (which is fixed by the mass of the central black hole).

One apparently easy way to explain these low luminosities is to postulate that the holes are accreting at very low rates. But the Galactic Centre says otherwise. The central light year is filled with a relatively thick plasma of gas emitted as winds from luminous stars, and from the known wind density one can calculate that the accretion rate onto the central mass has to be at least 10^{-4} solar masses a year. At 10% conversion efficiency, this would allow the hole to radiate at about 0.2% of the maximum Eddington rate. So unless the current accretion rate onto the central compact source happens to be orders of magnitude smaller than the above value⁵, the radiation efficiency of SgrA* must be less than a few times 10^{-6} , four to five orders of magnitude smaller than in standard accretion disk models.

One solution to this dilemma was proposed by Melia^{6,7}, who conjectured that the inflowing wind could have very low angular

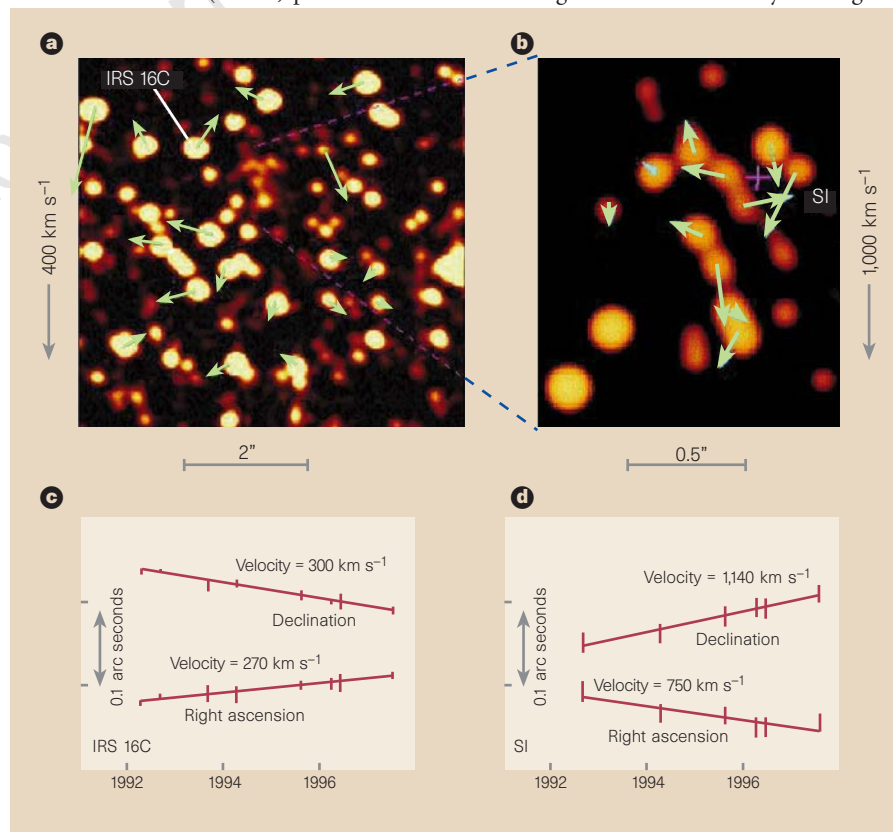


Figure 1 Stellar proper motions in the Galactic Centre. a, The derived proper motions in the central 6'' as vectors (green), with length proportional to the absolute value of the motions, superimposed on an infrared (2 μ m) image. b, The proper motion (blue) near the compact radio source SgrA* (cross). c, d, The change in right ascension and declination of two particular stars (marked on the upper images) between 1992 and 1997. The stellar motions systematically increase with decreasing distance from SgrA*, to values as large as 1,500 km s⁻¹ (assuming that the Galactic Centre is 8,000 parsecs away), requiring a compact, dark body of 2.6 million solar masses to exist within five light days of the radio source. Such a massive black hole would be expected to shine brightly in the infrared and X-ray bands, but its dimness may now be explained by the decoupling of electrons and ions in the accreting gas. (Figure adapted from ref. 4.)

momentum, and so flow in radially. Such purely radial 'Bondi-Hoyle' flows have very low radiation efficiencies. Melia then showed that a 10^{-4} -solar-mass yr^{-1} Bondi-Hoyle flow onto a million-solar-mass black hole would produce something like the observed radio to X-ray spectrum of SgrA*, including its weak infrared emission⁴. However, even small deviations from the purely radial flow would create a very large amount of infrared radiation, inconsistent with the observations; and such deviations would be likely to occur, at least transiently, through incomplete mixing of the incoming wind gas, or interactions with a pre-existing fossil accretion disk.

The new work of Narayan *et al.*¹ follows a second route. In a hot accretion flow, the gas is ionized to form a plasma. The heavy ions carry most of the mass, and thus of the energy, whereas the electrons produce most of the radiation (through synchrotron, Compton and Bremsstrahlung radiation processes). But, crucially, in a low-density flow the temperatures of the ions and of the electrons may decouple⁸⁻¹⁰. As a consequence, most of the gravitational energy would be viscously converted into thermal energy of the ions (in a hot ion torus^{9,10}), and not radiated away by the electrons.

Instead, the gravitational energy is carried ('advected') with the flow across the event horizon of the black hole^{11,12}. Such a flow leads to a low radiation efficiency even in a highly dissipative accretion disk. The possible application of such a model to the Galactic Centre had already been proposed more than a decade ago¹³, but Narayan *et al.* have now shown that a quantitative and self-consistent dynamical model can be constructed that fits most of the SgrA* observations for a 2.6-million-solar-mass black hole accreting at 10^{-4} solar masses per year. Like Melia's Bondi-Hoyle model, this advection-dominated accretion flow model also explains the spectral shape of the observed emission, with peaks in the millimetre and X-ray bands but very weak emission in the near-infrared and visible.

Turning the argument around, Narayan *et al.* point out that the success of their model (and the Melia model) leaves little doubt that the dark mass in the Galactic Centre must indeed be a black hole in the strict sense of General Relativity. Only if an event horizon exists can the gravitational energy truly disappear from sight. Otherwise there would have to be a 'surface' where the gravitational and thermal energy of the flow was converted to radiation after all. This argument adds to the already strong case for the existence of black holes.

Naturally, there remain some uncertainties. First, one might question how well the current accretion rate is determined by the observations. Apart from an order-of-magnitude uncertainty in the accretion rate

estimated above, time variability in the flow may be important. However, the dynamical time for the stellar winds near SgrA* is only about 100 years, and high-energy X-ray observations indicate that SgrA* was not very much more active 100 years ago than it is now. (Hard X-rays are scattered by dense, interstellar clouds. By looking at that scattered radiation from clouds about 100 light years from SgrA*, one can thus infer its X-ray brightness 100 years ago.) Second, the detailed plasma physics of the advection flow and the reality of the two-temperature solution are complex matters that depend on a number of poorly known parameters.

Nevertheless, the new work is the best answer yet to this strange paradox. Why are many massive black holes so black? It is because most of the time they are converting their food very inefficiently into radiation. □

Neurobiology

The analytical bend of the leech

L. F. Abbott

Descartes introduced the coordinate system that forms the basis of analytical geometry in an appendix to a philosophical discourse¹ that has been criticized for its contemptuous tone². Descartes might have been more humble had he known that his great discovery had already been implemented in the nervous systems of invertebrates, and that it is used by the leech to direct a reflexive motor behaviour. The evidence provided by Lewis and Kristan on page 76 of this issue³ arrives too late for Descartes, but in time to show the rest of us how a network of neurons computes a sensory-motor transformation.

Leeches respond to a touch on their skin by bending away from the point of contact. In each body segment, four sensory neurons (the P neurons) represent the location of the touch stimulus. They generate the response by activating a set of motor neurons through a network of interneurons. Lewis and Kristan have now determined the touch locations that evoke the maximum activity from each of the P neurons. Relative to the dorsal midline, these are at angles of 45, 135, 225 and 315 degrees, thereby forming two perpendicular axes (P₁ to P₄ in Fig. 1). When a touch is made at an angle θ , lying between the preferred angles of two P neurons, they respond by firing action potentials at rates proportional to $\cos(\theta)$ and $\sin(\theta)$, while the other two P neurons remain silent. Cosine and sine projections onto perpendicular coordinate axes are the defining features of a cartesian coordinate system. For any touch location, the firing rates of two active P neurons serve as cartesian coordinates. Four neurons are required to cover all directions because firing

Reinhard Genzel is at the Max-Planck Institut für Extraterrestrische Physik, W-8046 Garching, Germany. e-mail: genzel@mpe.mpg.de

- Narayan, R., Mahadevan, R., Grindlay, J. E., Popham, R. G. & Gammie, C. *Astrophys. J.* **492**, 554-568 (1998).
- Kormendy, J. & Richstone, D. *Annu. Rev. Astron. Astrophys.* **33**, 581-624 (1995).
- Miyoshi, M. *et al. Nature* **373**, 127-129 (1995).
- Genzel, R., Eckart, A., Ott, T. & Eisenhauer, F. *Mon. Not. R. Astron. Soc.* **291**, 219-234 (1997).
- Falcke, H., Biermann, P. L., Duschl, W. J. & Metzger, P. G. *Astron. Astrophys.* **270**, 102-106 (1993).
- Melia, F. *Astrophys. J.* **387**, L25-L28 (1992).
- Ozernoy, L. in *Back to the Galaxy* (eds Holt, S. S. *et al.*) 40-43 (AIP, New York, 1992).
- Ichimaru, S. *Astrophys. J.* **214**, 840-855 (1977).
- Shapiro, S. L., Lightman, A. P. & Eardley, D. M. *Astrophys. J.* **204**, 187-199 (1976).
- Rees, M. J., Begelman, M. C., Blandford, R. D. & Phinney, E. S. *Nature* **295**, 17-21 (1982).
- Begelman, M. C. *Mon. Not. R. Astron. Soc.* **184**, 53-67 (1978).
- Abramowicz, M., Czerny, B., Lasota, J. P. & Szuszkiewicz, E. *Astrophys. J.* **332**, 646-658 (1988).
- Rees, M. in *The Galactic Center* (eds Riegler, G. R. & Blandford, R. D.) *Am. Inst. Phys. Conf. Proc.* **83**, 166-176 (1982).

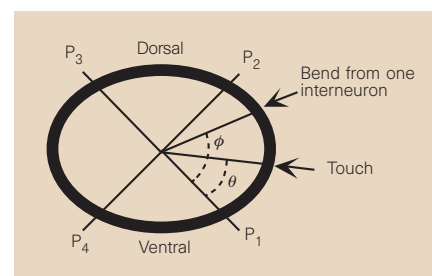


Figure 1 Computation of a sensory-motor transformation by neurons in the leech. The shaded area represents a cross-section of the body wall of the leech. Four sensory neurons (P cells) fire maximally in response to touches of the skin at the directions denoted by the lines P₁, P₂, P₃ and P₄. Lewis and Kristan³ have found that touches at other locations evoke responses in pairs of P cells at rates proportional to the cosine and sine of the angle θ defining the location of the touch. Also shown is the angle ϕ defining the bend produced by activation of a single interneuron.

rates cannot represent negative coordinate values. Here the leech is in good company — Descartes was also reluctant to use negative coordinates, and it was left to Newton to point out their usefulness².

As every student of introductory physics will testify, it is one thing to set up a cartesian coordinate system and quite another to use it computationally. Lewis and Kristan provide convincing evidence that the leech bending network takes full advantage of the cartesian representation it has established. They show that the touch location inferred by interpreting P-neuron fire rates as cartesian coordinates accurately matches the actual stimulus

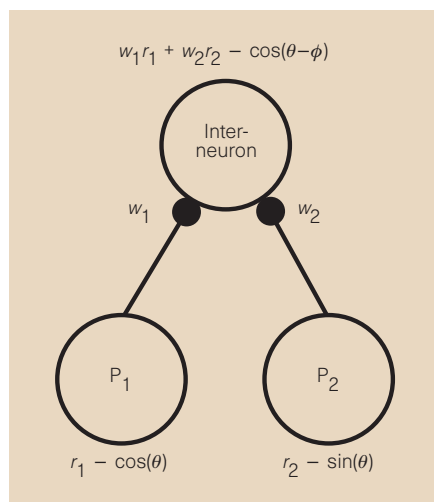


Figure 2 A simple circuit showing two P cells driving an interneuron. The synaptic drive to the interneuron is the sum of the synaptic weights w_1 and w_2 multiplied by the corresponding P-cell firing rates r_1 and r_2 . This drive is proportional to the cosine of the angle between the location of the touch and the bend produced by the interneuron.

location (3 per cent r.m.s. error). When they injected current into a P cell to modify its activity, the change in the bending response matched the shift in the coordinate value represented by the P cell's activity. Finally, they modelled how the sensory-motor transformation^{4,5} that directs the bend is computed.

The P cells activate motor neurons through a network that is estimated to contain 25 to 30 interneurons, 17 of which have been identified⁶. In the model, activation of an individual interneuron generates a bend in a particular direction characteristic of that neuron (the angle ϕ in Fig. 1). For an arbitrary touch location, different interneurons must be activated at appropriate levels so that they collectively produce the desired bend. A reasonable measure of the 'appropriate level of activation' is the projection of the desired bend direction onto the bend direction produced by the interneuron. If an interneuron generates a bend about the fixed angle ϕ when activated alone, this projection is proportional to $\cos(\theta - \phi)$. How does the bending circuitry compute this cosine?

Figure 2 shows the basic circuit element used in Lewis and Kristan's model. Two P cells drive an interneuron through synapses with strengths that are characterized by weights w_1 and w_2 (for simplicity the two inactive P cells are not shown). The total synaptic drive — given by the sum of the P-cell rates multiplied by the corresponding synaptic weights — is proportional to $w_1 \cos(\theta) + w_2 \sin(\theta)$. Any reader who dredges up the angle-addition law for cosines will realize that the desired quantity, $\cos(\theta - \phi)$, can be computed simply by making the synaptic weights w_1 and w_2 proportional to

$\cos(\phi)$ and $\sin(\phi)$. Lewis and Kristan found that the strengths of the synaptic connections from P cells to interneurons are, indeed, proportional to the appropriate cosines and sines for all 17 identified interneurons. Furthermore, a network model constructed on this principle duplicates the performance of the leech when at least 14 interneurons are included.

This is not the first time that a cartesian representation has been seen in an invertebrate sensory system. Miller, Jacobs and Theunissen⁷ found a virtually identical arrangement of four interneurons in the cricket cercal system, which responds to air currents. Neural firing rates that follow a cosine law have also been seen in a number of vertebrate systems, including the motor cortex of the monkey⁸, although not in an arrangement corresponding to a cartesian coordinate system. But until now we have not had a chance to see a nervous system making use of such a representation to do actual computations.

Nanotechnology

New tricks with nanotubes

M. S. Dresselhaus

Carbon nanotubes are seamlessly rolled single sheets of carbon atoms, only a few nanometres across. Two papers in this issue, Wildöer *et al.*¹ on page 59 and Odom *et al.*² on page 62, present strong experimental evidence for the remarkable electronic properties that were predicted³ in 1992 — namely, that they can be either metallic or semiconducting depending on their diameter and helicity. Both teams used scanning tunnelling microscope (STM) probes at low temperature to test these predictions, by measuring electronic structure (in terms of the density of states) and physical structure (the nanotube diameter and helicity). The results, obtained on nanotubes prepared differently and with different diameter distributions, are complementary.

The diameter and helicity of a nanotube are uniquely determined by the chiral vector (n, m) , where n and m are integers (Fig. 1). For example, a (10,0) nanotube simply has ten carbon hexagons around its diameter, and no helicity — this type is called a 'zigzag' tube, as a broken end perpendicular to the axis would have a zigzag shape. To go once around a (10,10) nanotube (a type of 'armchair'), one must move ten hexagons in one direction, turn 60° and move ten more; the helicity is 30°. Theory predicts that when $n - m$ is divisible by three, the single-walled carbon nanotubes are metallic; otherwise they are semiconducting. To verify this prediction, it is necessary to measure both the electronic properties and structure of individual nanotubes.

Although nanotube structures, resolved

Although the full circuitry of the bending network of the leech has not been worked out, it seems likely that one of its basic operating principles has been revealed. The P-cell/interneuron circuit is constructed to compute projections of various motor force vectors along a desired bend direction — and it performs this computation as elegantly as Descartes would have. □

L. F. Abbott is at the Volen Center and Department of Biology, Brandeis University, Waltham, Massachusetts 02254, USA.

e-mail: abbott@volen.brandeis.edu

1. Descartes, R. *La Géométrie* (Hermann, Paris, 1927).
2. Coolidge, J. L. *A History of Geometrical Methods* (Oxford Univ. Press, 1940).
3. Lewis, J. E. & Kristan, W. B. Jr *Nature* **391**, 76–79 (1998).
4. Salinas, E. & Abbott, L. F. *J. Neurosci.* **15**, 6461–6474 (1995).
5. Kristan, W. B. Jr, Lockery, S. R. & Lewis, J. E. *J. Neurobiol.* **27**, 380–389 (1995).
6. Lockery, S. R. & Kristan, W. B. Jr *J. Neurosci.* **10**, 1816–1829 (1990).
7. Miller, J. P., Jacobs, G. A. & Theunissen, F. E. *J. Neurophysiol.* **66**, 1680–1689 (1991).
8. Georgopoulos, A. P., Schwartz, A. B. & Kettner, R. E. *Science* **233**, 1416–1419 (1986).

on the atomic level by STM, have already been reported^{3,4}, the present papers collectively do a more systematic study. But more importantly, they relate the electronic properties of these nanotubes to their structure, primarily in terms of the one-dimensional (1D) density of electronic states along the tube axis. This is given by the derivative of the current–voltage curve, dI/dV , obtained by operating the probe as a scanning tunnelling spectroscopy (STS).

Theoretical predictions for the 1D density of electronic states (Fig. 2, overleaf) of both semiconducting (10,0) and metallic (9,0) nanotubes show a series of spikes. Each spike corresponds to the energy threshold for an electronic sub-band caused by the quantum confinement of electrons in the radial and circumferential directions of the nanotubes (which are one carbon atom thick and only

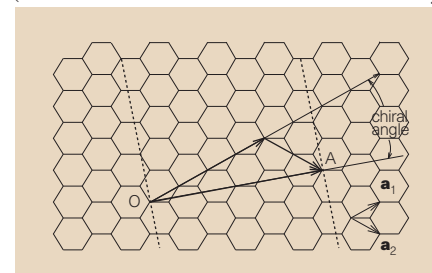


Figure 1 Nanotube coordinates. The vector OA defines a (4, 2) nanotube which, when laid out flat, would have edges along the dotted lines. To go around the tube once and get back to one's starting point, one moves four hexagons in the $\hat{a}_1$ direction and two hexagons in the $\hat{a}_2$ direction, 60° apart.

tens of atoms around). These spikes are very clearly seen in the STS studies¹ at 4 K. Furthermore, a non-zero density of states at the Fermi level E_F (the highest energy of electrons in a cold metal) is reported² in the metallic nanotubes, as expected³; and a vanishingly small density of states is seen in the semiconducting nanotubes. Further confirmation of the theory is provided by the much larger separation between spikes around E_F for metallic tubes relative to semiconducting tubes.

For semiconducting tubes, both groups also confirm an inverse linear dependence of the bandgap on the nanotube diameter, where the coefficient of proportionality is in good agreement with theoretical expectations.

But these papers^{1,2} also raise some worrying questions. The two groups find single-walled nanotubes with a wide range of helicities or chiral angles, in contrast to the narrow distribution of chiral angles (within 7.2° of the armchair 30° direction) found by electron diffraction studies⁵ — all using nanotubes prepared by the same group at Rice University^{1,6}.

Another puzzle concerns the 1D density of electronic states of chiral nanotubes, which generally have very large unit cells, often ten times larger than armchair or zigzag nanotubes of the same diameter. As the number of atoms in a unit cell increases, the number of spikes in the density of states should increase (over the range of energies encompassed by the π and π^* electron energy bands, which have the same total number of states per unit volume for all nanotubes). So armchair and zigzag nanotubes should have fewer but

sharper spikes than chiral nanotubes. But the new experiments¹ find about the same number of spikes for all semiconducting tubes of about the same diameter.

The headway made in determining the 1D density of states and (n, m) indices on the same tubes will lead to more quantitative and definitive studies on transport^{7,8} and optical properties of carbon nanotubes, especially in Raman spectroscopy⁹, where the vibrational spectra of specific nanotubes could be found by tuning the laser excitation frequency to be in resonance with the energy separation between spikes⁹.

Other promising new research directions include studying the modifications to the density of states induced by stress (by an STM tip, for example), by the nature of the substrate on which the nanotube sits, and by gates^{7,8} that introduce electric fields. And it will be vital to measure the 1D density of states for nanotube junctions¹⁰, bends, T-shapes¹¹ and caps¹², which are of particular

interest for device applications¹³, that might include transistors and electron emitters of nanometre dimensions early in the twenty-first century. □

M. S. Dresselhaus is in the Department of Electrical Engineering and Computer Science, and the Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. e-mail: millie@mngm.mit.edu

- Wildöer, J. W. G., Venema, L. C., Rinzler, A. G., Smalley, R. E. & Dekker, C. *Nature* **391**, 59–62 (1998).
- Odom, T. W., Huang, J.-L., Kim, P. & Lieber, C. M. *Nature* **391**, 62–64 (1998).
- Dresselhaus, M. S., Dresselhaus, G. & Eklund, P. C. *Science of Fullerenes and Carbon Nanotubes* (Academic, New York, 1996).
- Olk, C. H. & Heremans, J. P. *J. Mater. Res.* **9**, 259–262 (1994).
- Cowley, J. M., Nikolaev, P., Thess, A. & Smalley, R. E. *Chem. Phys. Lett.* **265**, 379–384 (1997).
- Thess, A. *et al. Science* **273**, 483–487 (1996).
- Tans, S. J. *et al. Nature* **386**, 474 (1997).
- Bockrath, M. *et al. Science* **275**, 1922–1924 (1997).
- Rao, A. M. *et al. Science* **275**, 187–191 (1997).
- Lambin, Ph. *et al. Chem. Phys. Lett.* **245**, 85–89 (1995).
- Menon, M. & Srivastava, D. *Phys. Rev. Lett.* **79**, 4453–4456 (1997).
- Carroll, D. L. *et al. Phys. Rev. Lett.* **78**, 2811–2814 (1997).
- Dresselhaus, M. S. *Physics World* **9**, May, 18–19 (1996).

Apoptosis

An endonuclease at last

Andrew Wyllie

Cell death is seldom chaotic. Rather, it is a subtly orchestrated disassembly of critical molecular structures, permitting the doomed cells to vanish with minimum disruption to the surrounding tissue. This idea was suggested by the uniform sequence of structural changes observed in cells dying in many different circumstances — the process of apoptosis¹. That such death might be effected by a dedicated set of gene products became clear when developmental cell death in the nematode *Caenorhabditis elegans* was shown to require the activity of several genes for which, to this day, no other major function has been identified².

On pages 43 and 96 of this issue, two papers from Shigekazu Nagata's group describe a mechanism for disassembly during apoptosis of the most complex of all biological molecules, chromatin^{3,4}. Despite earlier premonitions to the contrary, they find that chromatin is cleaved by a nuclease that is activated exclusively in apoptosis. And it is already clear that at least some elements of this mechanism are highly conserved between species.

Cleavage of chromatin at internucleosomal sites was first reported in dying cells by the Czech radiobiologist Miroslav Skalka⁵. Soon after, the 'chromatin ladder' pattern of regularly sized DNA fragments was recognized as characteristic of apoptosis⁶. Several candidates for the nucleases responsible were reported, but their properties were diverse enough to preclude identity with each other. Indeed, the view was expressed that DNA cleavage, appearing after the irreversible commitment to death, might be effected by

recruitment of any available nuclease, rather than requiring an apoptosis-specific enzyme.

Meanwhile, another family of effector molecules attracted more attention — the cysteine-containing aspartate-specific proteases, or caspases. The *ced3* gene product, which is necessary for all developmentally regulated cell death in the nematode, is the prototype of a large family of mammalian caspases at least some of which act in an autocatalytic cascade. Caspases cut their substrate proteins at tetrapeptide sites with characteristic motifs⁷. Caspase 3, for example, recognizes DEVD sequences and, during apoptosis, it cleaves and inactivates several significant cellular proteins in the cytosol, nucleus and cytoskeleton. Caspase 3 itself is activated by signals initiated by lethal stimuli of many different types and is thus a key player in turning on the pleiotropic effector events of apoptosis.

Working with cytosolic extracts, Nagata's group had already identified a factor that can cleave chromatin at internucleosomal sites⁸. Now they show that this activity depends on two interacting molecules³. The first is a hitherto unknown nuclease (M_r 40,000) — a 343-amino-acid, basic protein with a tell-tale nuclear localization signal at the carboxy terminus. The nuclease is expressed in many types of cell, but it is usually located in the cytoplasm in an inactive, latent form. In fractionated cytoplasmic extracts, it can be activated by caspase-3 digestion, earning it the name caspase-activated DNase (CAD).

The second molecule of the partnership is a strongly acidic protein (M_r 29,000–30,000; there are two isoforms of 265 and 331 amino

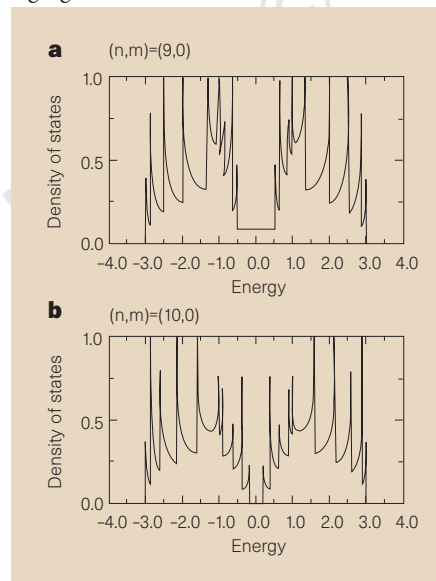


Figure 2 Calculated one-dimensional electronic density of states³ for a (9,0) nanotube (a) and a (10,0) nanotube (b). The (9,0) nanotube is metallic, but the bandgap near zero energy in b makes the (10,0) tube a semiconductor — so there is a huge difference in electronic properties from a small difference in geometry. These predicted properties have now been confirmed in measurements on single-wall nanotubes.

acids). This protein binds to, stabilizes and inactivates CAD, so it is called inhibitor of CAD (ICAD). ICAD releases CAD after caspase-3 digestion, and it is ICAD, not CAD, that possesses caspase cleavage sites. Using site-directed mutagenesis, Sakahira *et al.*⁴ show that one of these cleavage sites is essential for CAD release and activation. The authors also find that hyperexpression of ICAD in human Jurkat cells (a well-established test system for apoptotic stimuli) blocks the chromatin changes of apoptosis. Significantly, however, ICAD hyperexpression does not abrogate cell death, as shown by endogenous activation of caspase-3 activity, the characteristic surface-membrane expression of phosphatidylserine and eventual loss of mitochondrial oxidative function.

In these experiments, CAD and ICAD were purified from mouse cells, but the sequence of murine ICAD is closely homologous to a human protein, DFF45, discovered last year⁹ which, like ICAD, contains caspase 3 sites and is cleaved in apoptosis. Cleavage of DFF45 also permits nuclear endonuclease activation. The new data thus unequivocally show an apoptosis-specific pathway, initiated by caspase cleavage, which terminates in activation of a nuclease responsible for internucleosomal digestion of DNA. Interestingly, although the authors do not comment on this, their gels also show that active CAD cleaves chromatin to generate transient large fragments of uniform size before the appearance of the 'nucleosomal ladder': previous studies had indicated that the enzymes effecting these two functions might be different.

Why should there be a mechanism that specifically destroys DNA during apoptosis, and what pressures promote its development and expression? In apoptosis, caspase cleavage of DNA protein kinase and poly(ADP-ribose) polymerase unhook DNA repair from DNA damage^{10,11}. Lamin cleavage unpins the nuclear envelope¹². Cleavage of gelsolin, actin and intermediate filaments probably effect the hugely altered cell shape and movement of apoptosis¹³⁻¹⁵. So what does CAD do, and why should it be there?

There is little evidence that apoptosis-associated nuclease activity is sequence specific. But it can be argued that unrestricted DNA cleavage is essential for the successful completion of apoptosis, because unpackaged DNA from even one cell, uncoiling over some 1.5 m, could create an unmanageable extracellular environment for its neighbours. Further, DNA is dangerous stuff. Bacteria have restriction nucleases that destroy undesirable alien genomes, but a somewhat different strategy is needed to cope with extraneous DNA in eukaryotic tissues. The enormous size of the eukaryotic genome, and even the relative complexity of most viral DNA, make it improbable that cutting-site motifs could be found that satisfactorily discriminate viral from host DNA. Moreover, it is part

of the strategy of apoptosis that the entire nuclear DNA of the dying cell finds itself within neighbouring viable cells, and, in theory, this could colonize and scramble their genomes. What is needed is a mechanism that destroys DNA effectively (host as well as viral) in the process of removing virally infected or otherwise unwanted cells: in other words an endonuclease restricted to the process of apoptosis but not to any particular target sequence. And this is precisely what the CAD/ICAD pathway provides. The strategy of maintaining metazoan life seems to involve investment in a superbly designed, highly efficient and tightly controlled removal kit that is competent not only to kill unwanted cells on cue, but to bury the evidence, and fast. □

Andrew Wyllie is at the Sir Alastair Currie CRC Laboratories, Department of Pathology, University of Edinburgh, Edinburgh EH4 2XU, UK.

1. Kerr, J. F. R. *et al.* *Br. J. Cancer* **26**, 239-257 (1972).
2. Horvitz, H. R., Ellis, H. M. & Sternberg, P. N. *Neurosci. Comm.* **1**, 56-65 (1982).
3. Enari, M. *et al.* *Nature* **391**, 43-50 (1998).
4. Sakahira, H., Enari, M. & Nagata, S. *Nature* **391**, 96-99 (1998).
5. Skalka, M. *et al.* *FEBS Lett.* **72**, 271-274 (1976).
6. Wyllie, A. H. *et al.* *Nature* **284**, 555-556 (1980).
7. Thornberry, N. A. *et al.* *J. Biol. Chem.* **272**, 17907-17911 (1997).
8. Enari, M. *et al.* *Nature* **380**, 723-726 (1996).
9. Liu, X. *et al.* *Cell* **89**, 175-184 (1997).
10. Kaufmann, S. H. *et al.* *Cancer Res.* **53**, 3976-3985 (1993).
11. Casciola-Rosen, L. *et al.* *J. Exp. Med.* **183**, 1957-1964 (1996).
12. Takahashi, A. *et al.* *Proc. Natl Acad. Sci. USA* **93**, 8395-8400 (1996).
13. Kothakota, S. *et al.* *Science* **278**, 294-298 (1997).
14. Caulin, C. *et al.* *J. Cell Biol.* **138**, 1379-1394 (1997).
15. Na, S. *et al.* *J. Biol. Chem.* **271**, 11209-11213 (1996).

Nuclear transplantation

The science of the lambs



First Dolly, now Polly! In the 19 December issue of *Science* (278, 2130-2133; 1997), Angelika Schnieke and colleagues describe the generation of transgenic lambs using nuclear transfer — the technique that they developed to generate Dolly the sheep (Wilmut, I. *et al.* *Nature* **385**, 810-813; 1997).

But that's only half the story. The authors have created six lambs, including one called Polly (pictured), that should be able to produce milk containing factor IX — the blood-clotting protein that is deficient in people with haemophilia B. Not only do these results highlight the commercial potential of nuclear transfer, but they pave the way for its use in understanding other diseases.

Dolly was created by fusing the DNA of an adult udder cell with an egg from which the DNA had been removed. To make Polly and her sisters, Schnieke *et al.* used fetal fibroblast cells containing a transgene designed to express factor IX in the milk of the sheep and/or a selectable marker. After fusion with this DNA, the eggs were

implanted in surrogate Scottish Blackface ewes. Three of the resulting lambs contained the marker only, but the other three also carried the transgene.

Until now, the only way to produce transgenic animals has been by pronuclear injection. Between 200 and 300 copies of a transgene are introduced into a just-fertilized egg, and this is then implanted into a surrogate mother. But only around two per cent of the animals express the gene, and just a fraction of these produce commercially viable amounts of protein.

When Polly begins to produce factor IX, large quantities should be obtainable at low cost, and the protein will be free from the risks associated with purification from human plasma. Yields will be restricted only by the other (sheep) proteins in the milk, and a next step will be to try and knock out the genes that code for these. So Polly could be the first of many lines of transgenic sheep — and, as with Dolly, the crowds will undoubtedly flock to see her.

Alison Mitchell

Animal welfare

The physiology of the hunted deer

Georgia Mason

In the summer of 1997 one of Britain's largest landowners, the National Trust, banned the hunting of deer with hounds on its land. The decision, which caused a wave of protest from hunt-supporters, was based on a two-year study by Patrick Bateson and Elizabeth Bradshaw of how hunting affects the biology of red deer.

Late last month, the physiological data from the study were published¹, allowing scrutiny of the science behind the debate. They show that the effects of extended pursuits are severe — disruption of muscle tissue, the exhaustion of glycogen reserves, maximally high levels of cortisol and the breakdown of red blood cells. That being hunted with dogs causes stress is scarcely surprising. But these data reveal just how poorly adapted red deer are to predation by sustained pursuit; they also come at a time when moves to outlaw fox-hunting throughout Britain are gathering momentum.

Few would argue against the view that red deer populations need managing. But culling deer by hunting them with hounds, the hunters sometimes being mounted, sometimes on foot, is highly controversial (the deer are chased, and when captured are shot). Bateson's remit was to evaluate scientifically the suffering this causes². But what data can be used to address the question? Strictly speaking, the 'other minds' problem makes it impossible to tell the feelings of another animal, even a fellow human being³. But human emotions are usually accompanied by changes in behaviour and physiology — raised heart rates, sweaty palms, wincing, and so on — and it is this sort of correlate that allows inferences about the mental states of other species to be made⁴.

In practical terms, two complementary approaches are taken. One is to compare the behaviour and physiology of the animals under study with animals exposed to hunger, electric shocks, illness, injury and the like — experiences that are assumed to be unpleasant or already deemed unacceptable by society.

The other is to use the biological changes that occur in humans feeling fear, anxiety or pain, and to look for them in animals. Such changes include depressed immunity, gastric ulceration and altered functioning of the hypothalamus–pituitary–adrenal system, which is responsible for secreting corticosteroid hormones (cortisol, for example).

Bateson and Bradshaw¹ collected blood and muscle samples from 64 deer hunted over the course of a year. They also sampled 50 deer that had been stalked and cleanly shot with rifles. And in a part of the study not published here, they examined eight seriously injured deer which had, for instance, been hit by cars². These provided cases of suffering deemed 'unacceptable', because deer found in this state are generally euthanized as quickly as possible.

The physiological differences between hunted deer and those that were simply shot proved to be enormous. Not surprisingly, one difference was in the mobilization of fuel needed for physical exercise. For example, all hunted deer showed a rise in plasma glucose, followed by a fall to below baseline levels as stored carbohydrates were depleted. Muscle samples further revealed the extent of carbohydrate depletion: acidity levels declined the longer the hunt had gone on, showing that glycogen (which after death is converted to lactic acid⁵) had similarly declined. Indeed, in leg muscles, asymptotic levels of alkalinity were reached in any hunt that lasted longer than three hours.

From these data, the authors argue that, in at least a quarter of hunts that lead to a kill, carbohydrate reserves are at an absolute minimum, and that this is responsible for the animals' state of seeming exhaustion when they are captured. Hunted deer also had levels of cortisol 70 times those of shot animals, along with tenfold increases in both haem, the pigment released by damaged red blood cells, and two enzymes ordinarily at high levels only in striated muscle: creatine kinase and a form of lactate dehydrogenase.

So what does all this reveal about welfare? The interpretation of high cortisol concentrations, which are commonly taken to indicate stress, is complicated by the hormone's biological function (promotion of fuel mobilization when large amounts are needed⁶). But cortisol release during exercise is rarely maximal unless psychological stress is also involved⁶. The values reported here resemble the maximum seen in studies of the response of red deer to experimental administration of adrenocorticotrophic hormone. They are also higher than those seen in deer being transported by lorry for slaughter at abattoirs^{7,8}, and similar to those in the badly injured deer².

The release of creatine kinase from leaky myofibrils is common in humans doing physical exercise, but high levels can be taken to be a manifestation of actual pathology. In this study, 6.6 per cent of the deer had kinase levels greater than those indicating muscle damage in exercising horses⁹. High levels of this and lactate dehydrogenase are also symptomatic of a state called 'capture myopathy', in which acutely stressed animals become reluctant to move, hyperthermic and may die^{10,11}. Bateson and Bradshaw did not find the extreme acidosis often reported for wild ungulates (zebra and wildebeest, for example) which die after capture¹¹. But from the high levels of myofibril enzyme leakage, and the escape of haem from red blood cells (which might reflect hyperthermia), and from studies of white-tailed deer¹², the authors argue that some 8 per cent of red deer that escape the hunt may later die^{1,2}.

Apart from what it tells us about hunted deer, Bateson and Bradshaw's paper provides a model for investigations into the welfare consequences of other traditional British country sports, such as fox-hunting and hare-coursing (and, more internationally, big-game fishing, bull-fighting and bull-running). The stress physiology of any scientist bold enough to enter these highly emotional and political arenas would also make for a revealing study. □

Georgia Mason is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

1. Bateson, P. & Bradshaw, E. L. *Proc. R. Soc. Lond. B* **264**, 1–8 (1997).
2. Bateson, P. *The Behavioural and Physiological Effects of Culling Red Deer* (National Trust, London, 1997).
3. Warburton, N. *Philosophy — The Basics* (Routledge, London, 1992).
4. Mason, G. J. & Mendl, M. T. *Anim. Welf.* **2**, 301–320 (1993).
5. Warriss, P. D., Bevis, E. A. & Ekins, P. J. *Br. Vet. J.* **145**, 378–383 (1989).
6. Mason, J. W. J. *Psychiat. Res.* **8**, 323–333 (1971).
7. Goddard, P. J. *et al. Can. J. Zool.* **72**, 1872–1880 (1994).
8. Carragher, J. F., Ingram, J. R. & Matthews, L. R. *Appl. Anim. Behav. Sci.* **51**, 143–158 (1997).
9. Volfinger, L., Lassourd, V., Michaux, J. M., Braun, J. P. & Toutain, P. L. *Am. J. Physiol.* **266**, R434–R441 (1994).
10. McAllum, H. J. F. *R. Soc. N. Z. Bull.* **22**, 65–72 (1985).
11. Harthoorn, A. M., Van der Walt, K. & Young, E. *Nature* **247**, 577 (1974).
12. Beringer, J., Hansen, L. P., Wilding, W., Fischer, J. & Schiff, S. L. *J. Wildl. Mgmt* **60**, 373–380 (1996).



Figure 1 Hounded — a stag hunt (1839) painted by W. A. Kobell.

Cosmology

Density and destiny

David Branch

Will the Universe perpetually expand, or eventually collapse? Is space flat, or curved? On page 51 of this issue, Perlmutter *et al.*¹ report observations of the explosion of a star so distant that its demise preceded the birth of our planet — observations that begin to answer these grand questions, and promise a final answer soon.

If we inhabit a nice, simple Universe, the destiny of the Universe is determined by its density. The present expansion is being decelerated by gravity, at a rate governed only by the average density of matter, Ω_M . If $\Omega_M < 1$, space is negatively curved (like a saddle), and will expand for ever. $\Omega_M > 1$ means positive curvature (like a sphere) and eventual collapse. $\Omega_M = 1$ is supreme simplicity: space is flat and expansion slows asymptotically towards zero.

But empty space may have an energy density of its own, measured by the cosmological constant², Ω_Λ . Introducing this complicates matters, but there are reasons for doing so. Until recently, the main one was the 'cosmic age problem' — the expansion time of the Universe appears to be shorter than the ages of the oldest stars. The bigger Ω_M , the faster galaxies were separating in the past and the greater the problem (Fig. 1). But a positive Ω_Λ acts like a long-range repulsion which drives acceleration and increases our estimates of the expansion time. The age problem has eased in the past few years, however, as estimates of the Hubble constant (the present expansion rate) and the ages of the oldest stars have come down.

A more roundabout reason for introducing the cosmological constant is the theory of 'inflation'. According to the cosmic inflation hypothesis, the nascent Universe underwent such a phenomenal stretching that it should be nearly flat. But most observers, when they add the luminous matter to the dark matter inferred from galaxy motions, find $\Omega_M \sim 0.2$. This can be reconciled with inflation if $\Omega_M +$

$\Omega_\Lambda = 1$, but in that case Ω_Λ will rise towards one in the future as Ω_M plummets towards zero, and runaway acceleration will lead to another episode of inflation billions of years from now. Why should we just happen to be here at a time when Ω_M and Ω_Λ are about the same size?

When doing the Ω_M sum, however, it is hard to be sure that all matter is being counted; and Ω_Λ is even more elusive. A good way to measure both would be to observe the past history of the expansion, by looking deep into space, deep into the past. This boils down to plotting a Hubble diagram (apparent brightness against redshift) for a sample of 'standard candles' — things that have nearly the same true brightness, so relative brightness indicates relative distance. This is where the brightest kind of supernova comes in.

Type Ia supernovae³ (SNe Ia) are thought to be such standard candles. They start as white dwarf stars accreting matter from binary companion stars; eventually they approach a critical mass at which nuclear carbon burning begins. But under the electron-degenerate conditions in a white dwarf, this burning happens very rapidly — and the star explodes. These explosions all have about the same brightness because the objects involved are nearly identical: all made of carbon and oxygen, and all of the critical mass.

Indeed, the peak brightnesses of normal SNe Ia are so similar that treating them as standard candles gives relative distances to within 15 per cent. Correlations between peak brightness and distance-independent observables such as colour or the rate of fading can be used to standardize the candles further, and get distances accurate to 10 per cent. The absolute level of the SNe Ia peak brightness is needed for measuring the Hubble constant⁴, but not for Ω_M and Ω_Λ . The challenge there is to discover and accurately measure the apparent brightnesses of many



100 YEARS AGO

"Do the crystalline gneisses represent portions of the original earth's crust?" is the question asked, and answered in the affirmative, by Mr. J. Lomas, in his recent presidential address to the Liverpool Geological Society ... Excluding gneisses of later igneous or metamorphic origin, there remain the great series of fundamental gneisses, world-wide in distribution and uniform in general character, which must have had some world-wide cause of origin. As a possible cause of their foliation, Mr. Lomas suggests tidal action in the incompletely-consolidated crust. Prof. G. H. Darwin has shown that huge tidal wrinkles must have been raised by the moon when near the earth, forming ridges and troughs which ran north and south near the Equator, and curved to the eastward as they approached the Poles. The strike of the gneisses of Britain and Scandinavia corresponds to the direction of these tidal wrinkles in those latitudes, and there is evidence that the Palaeozoic strata were deposited in troughs parallel to the gneissic ridges.

From *Nature* 30 December 1897.

50 YEARS AGO

In a previous communication, a preliminary account was given of observations and experiments on the relative edibility of the flesh of birds. During the past two years, this work has been extended to an investigation of the relative edibility of birds' eggs. ... Each sample was tested in the form of a scramble, prepared over steam, a numerical score being awarded on a scale ranging from 10.0 (excellent flavour) to 2.0 (inedible). Eggs of eighty-one species have now been examined. In the following list these are arranged in order of palatability:

Domestic fowl	8.8
Coot (<i>Fulica a. atra</i> Linn.)	8.3
Moorhen (<i>Gallinula c. chloropus</i> (Linn.))	8.3
...	
Great tit (<i>Parus major newtoni</i> Prazak)	3.5
Blue tit (<i>Parus coeruleus obscurus</i> Prazak)	3.3
Wren (<i>Troglodytes t. troglodytes</i> (Linn.))	2.0

An extensive series of experiments has also been carried out to test the preferences of small egg-eating mammals such as the ferret, rat and hedgehog.

From *Nature* 3 January 1948.

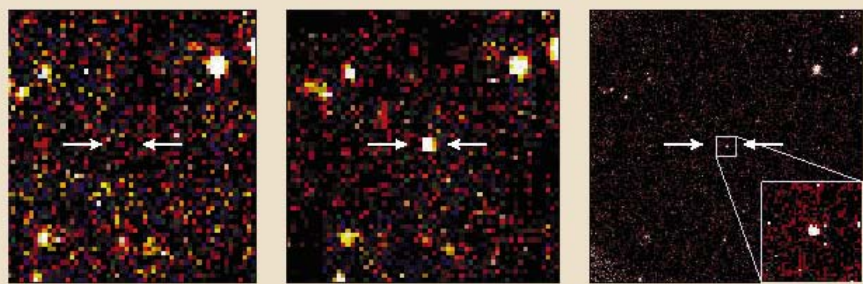


Figure 1 The furthest supernova. The first two images, from the Cerro Tololo Interamerican Observatory 4-metre telescope, show a galaxy halfway across the Universe transformed by the appearance of a type-Ia supernova (resolved more clearly in a follow-up observation using the Hubble Space Telescope). Being of well-known brightness, these supernovae help to determine the deceleration, and so the eventual fate, of the Universe.

very remote (high-redshift) SNe Ia.

Perlmutter's team has met this challenge by developing a search strategy that virtually guarantees the discovery of batches of high-redshift ($z > 0.3$) SNe Ia during an observing run at a large ground-based telescope⁵. The guarantee means that coordinated observations by other telescopes can be arranged in advance.

In this issue, Perlmutter *et al.*¹ present observations of SN1997ap, at the very high redshift $z = 0.83$. (The Universe has expanded since then by a factor of 1.83.) The brightness against time of SN1997ap was measured from the ground and from the Hubble Space Telescope, and the Keck telescope contributed a good spectrum to prove that it was a normal type Ia. The data do not support $\Omega_M = 1$. A rival pack of high-redshift-supernova hunters⁶ has also tentatively concluded that Ω_M is low. So if space is flat, there must be a cosmological constant.

The prospects for more precisely measuring Ω_M and Ω_Λ are bright. Both groups already have data on dozens of remote SNe Ia, including a number near $z = 1$; these will provide the leverage needed to determine Ω_M and Ω_Λ separately⁷. Astronomers will be wondering whether these analyses could be underestimating Ω_M . The inferred value

would come out higher if remote SNe Ia were observed to be brighter than they are, relative to low-redshift supernovae. Could the high-redshift SNe Ia of the younger Universe be intrinsically dimmer than the nearer ones? (Not likely, if they have normal spectra.) Could there be wavelength-neutral intergalactic obscuration? Could the low-redshift sample be seriously biased by observational selection effects?

If the answers to these and other such questions prove to be no, it will become clear that expansion is here to stay. We should keep in mind, though, that if the cosmological principle (large-scale homogeneity) isn't valid, we could simply be exploring the nature and future of just one bubble in a cosmic sea of champagne. □

David Branch is in the Department of Physics and Astronomy, University of Oklahoma, Norman, Oklahoma 73019, USA.

e-mail: branch@phyast.nhn.yoknar.edu

1. Perlmutter, S. *et al.* *Nature* **391**, 51–54 (1998).
2. Carroll, S. M., Press, W. H. & Turner, E. L. *Annu. Rev. Astron. Astrophys.* **30**, 499–542 (1992).
3. Ruiz-Lapuente, P., Canal, R. & Isern, J. (eds) *Thermonuclear Supernovae* (Kluwer, Dordrecht, 1997).
4. Saha, A. *et al.* *Astrophys. J.* **486**, 1–20 (1997).
5. Perlmutter, S. *et al.* *Astrophys. J.* **483**, 565–581 (1997).
6. Garnavich, P. *et al.* *Astrophys. J.* (in the press).
7. Goobar, A. & Perlmutter, S. *Astrophys. J.* **450**, 14–18 (1995).

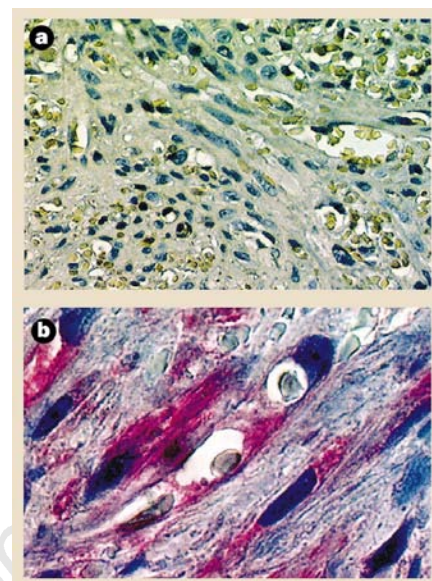


Figure 1 Kaposi's sarcoma biopsy. a, Numerous vascular spaces and extravasated red blood cells are surrounded by spindle ('tumour') cells. b, Higher magnification of spindle cells, showing abundant magenta staining for the powerful angiogenic protein, vascular endothelial growth factor. (Reproduced, with permission, from ref. 2.)

Kaposi's sarcoma

Coupling herpesvirus to angiogenesis

Chris Boshoff

Since the discovery of Kaposi's sarcoma-associated herpesvirus (KSHV) only three years ago¹, and its detection in every Kaposi's sarcoma biopsy (Fig. 1), there have been many hypotheses to explain how this virus contributes to the abundant vasculature and spindle-cell proliferation that are associated with the disease. Spindle cells are considered to be the 'tumour' cells, and they contain (and secrete) large amounts of a pow-

erful angiogenic agent called vascular endothelial growth factor (VEGF)². Angiogenesis — the sprouting of new blood vessels from pre-existing ones — is essential for tumour progression and, on page 86 of this issue, Bais *et al.*³ report that the G-protein-coupled receptor (GPCR) encoded by KSHV induces an 'angiogenic switch' in transformed NIH3T3 cells.

Like experimentally mutated cellular

GPCRs, KSHV-GPCR is fully active in the absence of chemokine ligands⁴. Bais *et al.*³ show that KSHV-GPCR can, indeed, act as a viral oncogene to transform NIH3T3 cells, and that this transformation is accompanied by secretion of VEGF. Spindle cells in Kaposi's sarcoma belong to the endothelial lineage, but they often express antigens characteristic of endothelial, macrophage and smooth-muscle cells. This indicates that they either represent pluripotent mesenchymal precursors, or are endothelial cells undergoing aberrant differentiation. KSHV DNA is present in most spindle cells, and also in endothelial cells and monocytes in Kaposi's sarcoma lesions.

The KSHV-GPCR gene is not the only viral oncogene to induce tumour formation and angiogenesis. Cells transformed by v-Ha-ras and v-raf (the constitutively active retroviral forms of cellular *ras* and *raf*) express increased

Viral pirates on a cellular sea

KSHV-GPCR is just one of an extraordinary number of genes that are pirated^{8,9} from eukaryotic cellular DNA by KSHV. The structural proteins and viral enzymes that are common to most herpesviruses probably originated from an ancient progenitor of contemporary herpesviruses. But the recognizable cellular genes occur only sporadically in some herpesviruses. They

are probably more recent acquisitions from the host genome, and they might support viral replication in a specific microenvironment (which, for KSHV, could be the microvasculature).

The captured eukaryotic genes have acquired unique properties (through accelerated evolution within the viral genome) which can give us insight into the biology of their cellular

counterparts, and may even be exploited for development of therapies. For example, the D-type cyclin encoded by KSHV binds to its activating cellular partner cyclin-dependent kinase 6 (Cdk6). This association seems to be resistant to proteins that inhibit Cdk6 (ref. 11), meaning that the KSHV-cyclin D complex can drive cellular proliferation, and

that this is not blocked by the normal cell-cycle control mechanisms. Similarly, the chemokines that are encoded by KSHV are more promiscuous than their cellular counterparts¹². Moreover, the viral chemokines induce angiogenesis and might, therefore, also be directly involved in the pathogenesis of Kaposi's sarcoma¹².

C.B.

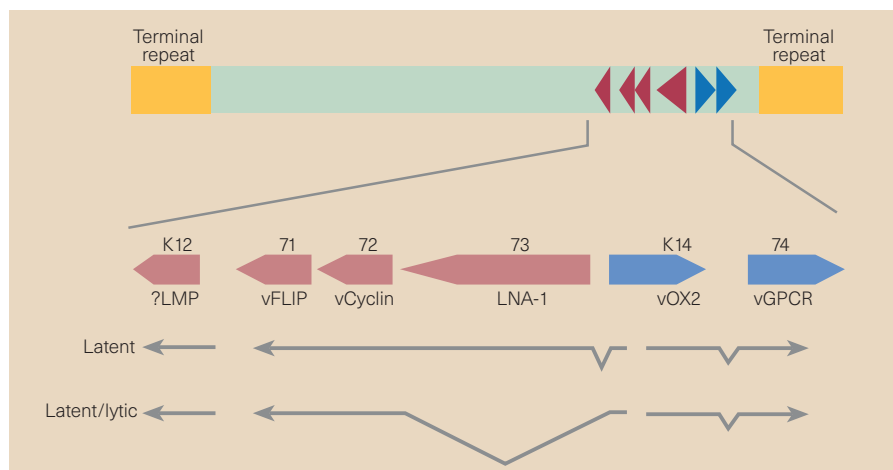


Figure 2 'Oncogenic cluster' in the genome of Kaposi's sarcoma-associated herpesvirus (KSHV). The G-protein-coupled receptor (GPCR) is encoded by a viral gene that is transcribed with a homologue of OX2. Four other viral genes are transcribed in the opposite direction, and from different promoters. Bais *et al.*³ have shown that GPCR acts as a viral oncogene, and that cellular transformation is accompanied by secretion of vascular endothelial growth factor. K12 is a latent membrane protein (LMP) that may act like one of the latent membrane proteins of Epstein-Barr virus.

VEGF⁵. But these genes induce a different signalling pathway from KSHV-GPCR. Bais *et al.* have found that GPCR triggers the JNK/SAPK and p38MAPK pathways, although they have not shown whether these cascades are involved in the induction of VEGF. So as many as three major signalling pathways that link membrane receptors with the nucleus in mammalian cells could, potentially, be involved in tumour progression by switching on angiogenesis. The viral interferon regulatory factor is another KSHV-encoded protein that can transform NIH3T3 cells and induce tumour formation in nude mice⁶.

The KSHV-encoded GPCR is part of a potential 'oncogenic cluster' of proteins within the viral genome. It is transcribed with OX2 (S. Talbot, unpublished data), which is a protein involved in intercellular signalling. This transcript is in the opposite direction — and from a different promoter — to three viral genes (Fig. 2), which encode cyclin D, vFLIP (viral FLICE inhibitory protein), and an immunogenic latent nuclear antigen, LNA-1. The vFLIP protein contains a death-effector domain, and it is homologous to the cellular FLIP protein⁷. Like cellular FLIP, vFLIP might block one of the main pathways by which immune mechanisms cause cell death — induction of apoptosis by the tumour-necrosis factor family of receptors. LNA-1 resembles the Epstein-Barr virus nuclear antigens (EBNAs). It may transactivate other cellular or viral genes to promote cellular growth or, like EBNA-1, it could maintain the viral DNA in an extrachromosomal circular (episomal) form.

Herpesviral proteins can be classified as being either 'latent' or 'lytic'. Latent proteins are expressed during non-replicative infection, whereas lytic proteins are expressed during viral replication, and they are associated with cell death. In KSHV, the expression pat-

tern of the pirated genes (see box) is more murky — many genes are expressed during both the lytic and latent phases, although they are upregulated during lytic infection⁸. The pattern of viral gene expression also seems to differ in haematopoietic compared with endothelial cells. For example, KSHV-encoded interleukin-6 is expressed in latently infected B cells, but not in spindle cells⁸.

KSHV-GPCR is expressed in latent B cells, and upregulated with lytic induction. Both types of KSHV infection are present in Kaposi's sarcoma⁹, and both probably contribute to the complex pathology of the dis-

ease. Most spindle cells in Kaposi's sarcoma are latently infected, although many might undergo lytic infection at some point. Lytic infection, with production of viral particles, might be necessary to drive the formation of lesions. It is also possible that 'abortive lytic' infection (meaning that most early lytic genes are expressed) contributes to the formation of tumours in Kaposi's sarcoma.

The complex histology and expression pattern of KSHV proteins in Kaposi's sarcoma indicate that the model of tumorigenesis might not be like any other virus-induced malignancy. The irony of current KSHV research is that, despite having all this ammunition to trigger a malignancy, nobody has yet shown that KSHV can transform any cell type *in vitro*. Moreover, only a fraction of immunocompetent infected patients will actually develop a tumour associated with this virus. □

Chris Boshoff is at the Chester Beatty Laboratories, Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK.

e-mail: cboshoff@icr.ac.uk

1. Chang, Y. *et al. Science* **266**, 1865–1869 (1994).
2. Cornali, E. *et al. Am. J. Pathol.* **149**, 1851–1869 (1996).
3. Bais, C. *et al. Nature* **391**, 86–89 (1998).
4. Arvanitakis, L., Geras-Raaka, E., Varma, A., Gershengorn, M. C. & Cesarman, E. *Nature* **385**, 347–349 (1997).
5. Grugel, S., Finkenzeller, G., Weindel, K., Barleon, B. & Marme, D. *J. Biol. Chem.* **270**, 25915–25919 (1995).
6. Gao, S.-J. *et al. Oncogene* **15**, 1979–1985 (1997).
7. Irmeler, M. *et al. Nature* **388**, 190–195 (1997).
8. Moore, P. S., Boshoff, C., Weiss, R. A. & Chang, Y. *Science* **274**, 1739–1744 (1996).
9. Orenstein, J. M. *et al. AIDS* **11**, 735–745 (1997).
10. Neipel, F., Albrecht, J.-C. & Fleckenstein, B. *J. Virol.* **71**, 4187–4192 (1997).
11. Swanton, C. *et al. Nature* **390**, 184–187 (1997).
12. Boshoff, C. *et al. Science* **278**, 290–293 (1997).

Low temperature physics

Condensate forced out of hiding

Allan Griffin

Below a temperature of 2.17 K, liquid ⁴He undergoes a phase transition to a new state of matter with superfluid properties. It has long been accepted that a Bose–Einstein condensate is responsible, namely that a macroscopic number of atoms occupy the same quantum state. Direct evidence for this condensate has eluded our best efforts for many decades, but finally, as reported on page 56 of this issue by Wyatt¹, a macroscopic fraction of atoms with zero momentum has been observed in liquid ⁴He. A Bose condensate of Cooper pairs also underlies superfluidity in superconductors and liquid ³He.

Wyatt used high-energy phonons generated in the bulk liquid to knock atoms from the free surface of superfluid ⁴He (Fig. 1, overleaf), and then measured their kinetic energy and momentum². In the new analysis¹, Wyatt finds that all the evaporated atoms initially had zero momentum parallel

to the liquid surface. These results end a long and sometimes controversial search in low-temperature physics, and development of the technique opens up the exciting possibility of producing a phase-coherent beam of ⁴He atoms, a kind of 'atom laser'.

Recalling a little history should make the significance of these results clear. Within weeks of the discovery of the superfluid properties of liquid ⁴He in 1938, Fritz London³ suggested that they were a consequence of the fact that ⁴He atoms have zero spin, and hence obey Bose statistics. (Bosons are particles with integer spin 0, 1, 2, ...; fermions have spin 1/2, 3/2, ...). In particular, London argued that superfluidity was related to the occurrence of Bose–Einstein condensation (BEC) in a liquid, similar to that predicted in an ideal gas a decade earlier by Einstein⁴ in a paper that had fallen into disfavour. London's colleague Tisza⁵ quickly developed a two-fluid model, in which the superfluid

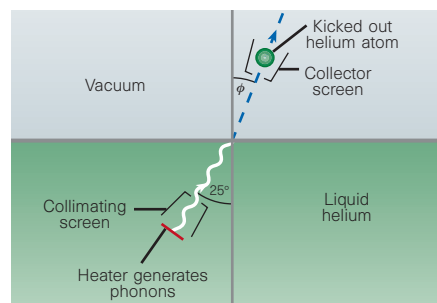


Figure 1 Quantum evaporation with liquid helium. Phonons from a heated metal film are collimated at 25° to the vertical, and kick atoms from the surface of a tank of superfluid ^4He . These atoms come off at an angle ϕ of around 20° to the vertical, implying that they start with zero momentum parallel to the surface — in other words, that they form a ground-state Bose–Einstein condensate. (Adapted from ref. 2.)

component had its origin in the macroscopic occupation of atoms of a single quantum state. The superfluid component is a new, coherent, collective degree of freedom which moves without resistance.

It took many years to properly develop this BEC picture of London and Tisza, and it was overshadowed by the phenomenological theory developed in 1941 by L. D. Landau, who derived the two-fluid equations of motion describing superfluidity without any explicit reference to a Bose condensate. However, by about 1960, it was generally agreed that the Landau theory could be understood as a consequence of a condensate described by a macroscopic wavefunction (for a review, see ref. 6). The superfluid velocity is given by the gradient of the phase of this condensate wavefunction.

The superfluid fraction can be easily measured, but in a Bose liquid it is not simply related to the underlying condensate (in contrast to a Bose gas), as shown by the fact that the entire liquid forms the superfluid at zero temperature. In spite of this, the condensate fraction and its temperature dependence have been accurately determined by indirect methods over the past two decades. Computer simulations⁷ show consistently that it is about 9% at absolute zero, and vanishes at the superfluid transition temperature of 2.17 K. That agrees with inelastic neutron-scattering experiments, which allow one to extract the atoms' momentum distribution. Although the condensate peak at zero momentum is masked by finite energy resolution and final-state effects, it is possible instead to find the number of atoms with finite momentum (the non-condensate fraction)^{6,8}, and thus obtain the condensate fraction.

Those studying liquid helium could only look on with envy, three years ago, when Bose condensation was clearly observed in alkali atoms in magnetic traps⁹. But now Wyatt's evaporation experiment has finally provided direct evidence for a condensate in liquid ^4He

at 100 mK, where the liquid is essentially in its ground state. The large number of atoms that emerge near the expected angle of 20° to the vertical (after absorbing mono-energetic phonons) means there must be macroscopic occupation of the state with zero momentum parallel to the liquid surface. In contrast, the flux of evaporated atoms initially in states with finite momentum leads to a broad angular spread of low intensity. Before we can estimate the condensate fraction, measurements of greater sensitivity are needed to detect the evaporation from the non-condensate atoms.

Crucial to the success of the new quantum evaporation experiment is the ability to produce a narrow mono-energetic beam of long-lived, high-energy (10.15 K) phonons. The remarkable physics of the anharmonic up-scattering phonon processes which gives rise to this beam was clarified only a few years ago¹⁰.

The observation that the evaporated atoms have absorbed a single high-frequency phonon implies that these atoms come from the region very close to the liquid surface, otherwise there would be energy losses via collisions before the atom evaporates. That limits Wyatt's technique to the study of ^4He atoms in the surface region — perhaps a blessing in disguise, as the zero-temperature condensate fraction is predicted to go from 9% in the bulk liquid to 100% in the low-density surface tail (due to lack of collisions)^{11,12}. Future evaporation experiments promise to give detailed information about the profile of the inhomogeneous Bose condensate at liquid ^4He surfaces.

Wyatt's experiment has put the Bose condensate in liquid ^4He back into the spotlight, 60 years after it was suggested by London³. As an agreeable side-benefit, this quantum evaporation technique might be adapted to produce a phase-coherent mono-energetic beam of ^4He atoms, as they all come from the same initial quantum state. This would be similar to the beam produced in 1997 by allowing a Bose-condensed gas of ^{23}Na atoms to leak out of a trap¹³ — another type of 'atom laser'. □
Allan Griffin is in the Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada.

- Wyatt, A. F. G. *Nature* **391**, 56–59 (1998).
- Brown, M. & Wyatt, A. F. G. *J. Phys. Cond. Matter* **2**, 5025–5046 (1990).
- London, F. *Nature* **141**, 643–644 (1938).
- Einstein, A. *Sitzungsber. Berlin Preuss. Akad. Wiss.* 3–14 (1925).
- Tisza, L. *Nature* **141**, 913 (1938).
- Griffin, A. *Excitations in a Bose-Condensed Liquid* (Cambridge Univ. Press, 1993).
- Ceperley, D. M. & Pollock, E. L. *Phys. Rev. Lett.* **56**, 351–354 (1986).
- Sokol, P. E. in *Bose–Einstein Condensation* (eds Griffin, A., Snoke, D. W. & Stringari, S.) 51–85 (Cambridge Univ. Press, 1995).
- Anderson, M. H. *et al. Science* **269**, 198–201 (1995).
- Tucker, M. A. H. & Wyatt, A. F. G. *J. Phys. Cond. Matter* **6**, 2813–2824; 2825–2834 (1994).
- Lewart, S. D., Pandharipande, V. R. & Pieper, S. C. *Phys. Rev. B* **37**, 4950–4964 (1988).
- Griffin, A. & Stringari, S. *Phys. Rev. Lett.* **76**, 259–263 (1996).
- Andrews, M. R. *et al. Science* **275**, 637–641 (1997).

Daedalus

The Brain-cleaner

The brain is a statistical computer. Its neurons are richly connected by synapses, each of which holds a probability that a pulse from one neuron will trigger the other. Everything we know, it is claimed, has been acquired and is stored as innumerable synaptic firing probabilities.

This cannot be the whole truth. Nerves fire in milliseconds; but it can take many weeks or months to learn a skill, see an argument or change our mind. We all waste years of life nursing unrealistic hopes, recovering from emotional setbacks, coming to terms with limitations or advancing age, or abandoning foolish compulsions or delusions. Why is mental change so slow? Daedalus points out that nerves are hollow; they are filled with fluid axoplasm. This contains not merely the ions of nerve action, but cell nutrients, enzymes, and so on. These diffuse slowly along the nerve, typically at a few millimetres a day. (The rabies virus makes its relentless, deadly passage from bite to brain by diffusing up the nerves at this sort of rate.) If the firing probability of a synapse can be altered only by the diffusion of some key component down the axon, the slow pace of human mental change would be explained.

Daedalus now plans to speed it up. A blow on the head is popularly supposed to shake one's ideas up; and shock could well help to stir the cerebral axoplasm. But ultrasonics seems more controllable. An ultrasonic cleaning bath works by very small-scale stirring indeed. DREADCO's cunning new 'Brain-cleaner' launches a highly asymmetric, ultrasonic waveform into the skull. Axoplasm has a non-newtonian viscosity. The sharp leading edge of each cycle shifts it abruptly one way, but the reduced shear of its gentler trailing edge does not reverse the shift exactly. So the axoplasm convects strongly, updating the synaptic probabilities in minutes rather than months.

Regular brain-cleaning will save us all years of life. Emotional traumas, from failed love-affairs to bereavements and sackings, will be swiftly put behind us. The economic cycle will speed up or even be damped out, as economists and business tycoons no longer lag behind the changing reality. Generals may at last cease to prepare for the previous war. Psychoanalysts will go bankrupt in droves: their patients will be transformed after mere days on the couch. Technological advance will become even more breathless, and the world of fashion will change with dizzying speed.

David Jones

Dali's dimensions

Salvador Dali's image of the crucified Christ represents the search for the transcendental or fourth dimension. Artists, theologians, mathematicians and cosmologists down the ages have all joined in the quest.

Martin Kemp

A recurrent theme of world art has been the striving to reach out into transcendental realms. Any artist so minded inevitably has to confront the paradox that transcendence implies access to dimensions (spatial, material, temporal, conceptual...) beyond those that are open to literal description by the solid surfaces of visible artefacts.

Early in this century, pioneers of modernism in France, Russia and other centres of the avant-garde seized in various ways upon the popularized notions of non-Euclidean geometry, the fourth dimension, infinity and relativity as providing grounds for the rejection of traditional illusions of three-dimensional space. The levels of understanding achieved by artists varied greatly, and the general picture is one of opportunistic annexing of new theories in a schematized manner.

One approach, particularly favoured by Russian artists such as Casimir Malevich and El Lissitzky, was to exploit abstraction as a way of negating conventional space. Malevich's famous *Black Square* and *White on White* use extreme reduction of means to evoke the infinity and nothingness that lie beyond material definition, but the effect relies on absence and denial rather than positive visualization.

The dilemma was not one exclusive to artists. All the ingenious attempts by mathematicians and cosmologists to provide visual tools for the concrete envisaging of the fourth dimension have struggled with the irredeemably three-dimensional parameters of the sensate space beyond our bodies and within our minds — and with the limited representational techniques available to us.

The kinds of diagrammatic solutions proposed at the end of the last century by the eccentric mathematician and convicted bigamist, Charles Howard Hinton, seem to be about the best we can do. His unfolded four-dimensional hypercube or tesseract — as the spatial counterpart of a normal cube that has been unfolded into a flat template — became a popular and mystical symbol of the transcendent spaces of the new mathematics.

As such it was adopted by the renowned Spanish Surrealist, Salvador Dali, in his *Corpus* (or *Christus*) *Hypercubus*, in which Christ's sacramental body is transfixed ambiguously within the foremost of the

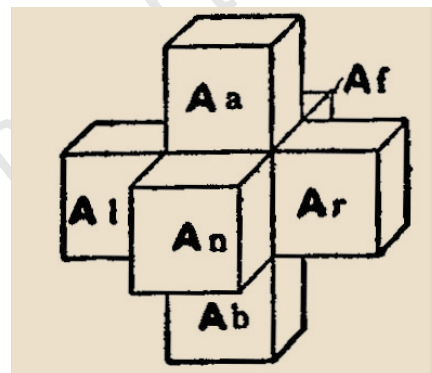
cubes of a Hintonian tesseract, floating in front of the Virgin above a pavement decorated with an unfolded 3D cube.

Is this more than slick visual opportunism or, at best, knowing symbolic allusion? We may gain some comfort from learning that Thomas Banchoff, the author of *Beyond the Third Dimension* (W. H. Freeman, 1990), had been contacted by the artist and was impressed by his "technical knowledge". Yet doubts remain. Dali was not working visually with anything beyond standard cues for the representation of forms in space.

However, in our present context, Dali's painting does stand effectively for an age-old striving in art, theology, mathematics and cosmology for access to those dimensions that lie beyond the visual and tactile scope of the finite spaces of up-and-down, left-and-right, and in-and-out that imprison our common-sense perceptions of our physical

world. The scientists' success in colonizing the extra dimensions is defined mathematically. The artists — and Dali's corporeal Christ — reach out by visual inference. □

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.



Salvador Dali's *Corpus Hypercubus* (Crucifixion), 1954, and (above) the hypercube of Charles Howard Hinton, from his book *The Fourth Dimension*, 1912.

METROPOLITAN MUSEUM OF ART (CHESTER DALE COLLECTION), NEW YORK/DACS, 1998

False teeth of the Roman world

The history of dental implants has been looked at before¹⁻⁴ but there have been no documented cases of a functional implant from ancient times. We report here a wrought iron dental implant of a right second upper premolar from a Gallo-Roman necropolis at Chantambre (Essonne, France), from the first or second century AD. The implant and the socket fit perfectly together and the osseointegration appears viable.

Since the 1960s the dental profession has been divided on the issue of dental implants⁵. Thanks to the osseointegration principle^{6,7}, ideas about suitable interfaces between the bone and the implant have made considerable progress, and it has been shown that it is not desirable to insert fibrous tissue⁸. The discovery of an osseointegrated implant from a Gallo-Roman population shows that direct contact between the implant and the bone can be obtained using surprisingly basic technology.

The individual in question, a man who was over 30 years old when he died, is dated to the end of the first century AD or the beginning of the second century by associated pottery and radiocarbon dating. There is a piece of metal where the right second upper premolar would have been (Fig. 1a) which has been severely corroded, preventing magnification for the study of the microstructure. The central part, however, has not been affected. X-ray microanalysis (energy dispersive spectroscopy), together with scanning electron microscopy of the apical fragment, identify it as iron or non-alloy steel. The main chemicals are iron and oxygen; nevertheless the presence within a particular section of dark zones — bearing traces of silicon (0.73%) and calcium (0.26%) — and of clear zones shows that some zones are more oxidized than others. This may imply that the metal was given its shape through a hot-hammering and folding process.

The retro-alveolar X-ray (Fig. 1b) shows that the alveolar wall and the pseudo-root

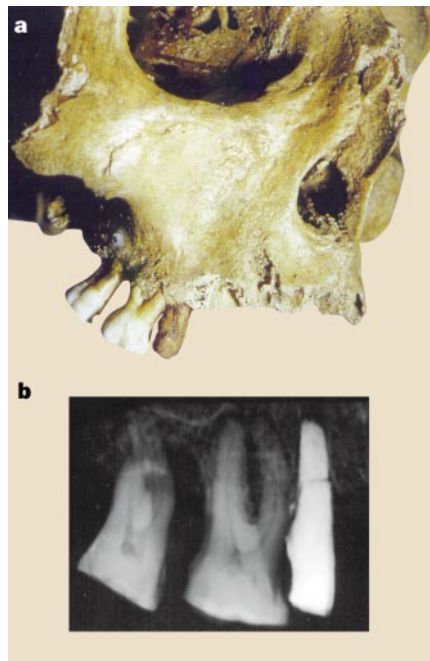


Figure 1 Details of the false tooth. **a**, Antero-lateral view of the skull. The iron-made dental implant is where the right second upper premolar would have been. **b**, Retro-alveolar X-ray of the right maxilla. The implant was dropped after being handled and it broke near its apical third into two pieces that were then glued together again. The line of breakage is visible on this X-ray picture. The alveolar wall and the pseudo-root fit perfectly together. Only an area one to two millimetres high in contact with the mesial alveolar wall shows an absence of contact between the bone and the implant.

fit perfectly together without peri-implant radiolucent areas, which is considered to be one of the success criteria for the implant of a dental prosthesis⁹. According to available data⁵, osseointegration requires a minimum of three to six months. The minor reactions of the periodontium indicate that the subject was probably fitted with his implant more than a year before he died. The osseointegration implies that the man who made the implant used the original tooth as

a model. The implant was probably set by impaction soon after the tooth loss. Chance played a part in this success, but the technical choices were, in retrospect, conducive to osseointegration. Furthermore, although iron is surely not the ideal metal for dental implants, its rugged surface must have provided satisfactory adhesion to the bone¹⁰.

Because it is osseointegrated and in a good anatomical position, this implant might have been functional. We cannot know why it was inserted. But the early disappearance of the left molars might have given the subject the desire to keep an active right side.

This case, in addition to its exceptional aspect and the technical craft it required, gives remarkable clues about medicine and anatomy in this rural community of the first or second century AD. This implant reflects the potential of early medicine and the validity of the osseointegration principle^{4,5}.

Eric Crubézy

Fédération d'Anthropologie,
Université Toulouse III,
31000 Toulouse, France

Pascal Murail, Louis Girard

Laboratoire d'Anthropologie,
URA 376 du CNRS, Avenue des facultés,
33405 Talence Cedex, France

Jean-Pierre Bernadou

Ecole Nationale Supérieure de l'Aéronautique
et de l'Espace,
31055 Toulouse Cedex, France

1. Ortner, D. J. *Paleopathol. Newsl.* **53**, 10–11 (1986).
2. Vallois, H. V. *L'Anthropologie* **75**, 191–220 (1971).
3. Granat, J. *L'Information Dentaire* **72**, 1959–1961 (1990).
4. Capasso, L. *Archeologia* **5**, 52–55 (1986).
5. Bert, M. & Missika, P. *Les Implants Ostéo-intégrables* (CDP, Paris, 1992).
6. Brånemark, P. I. et al. *Scand. J. Plast. Reconstr. Surg.* **3**, 81–91 (1969).
7. Brånemark, P. I. et al. *Scand. J. Plast. Reconstr. Surg.* **16** (suppl.), 1–132 (1977).
8. Bert, M. *Actualités Odonto-stomatologiques* **159**, 507–520 (1987).
9. Albrektsson, T., Zarb, G., Worthington, P. & Reiksson, A. R. *Int. J. Oral Maxillo-Fac. Impl.* **1**, 11–25 (1986).
10. Claes, L., Hutzschenreuter, P. & Pohler, O. *Arch. Orthop. Unfallchir* **85**, 155–163 (1976).

Breeding phenology and climate...

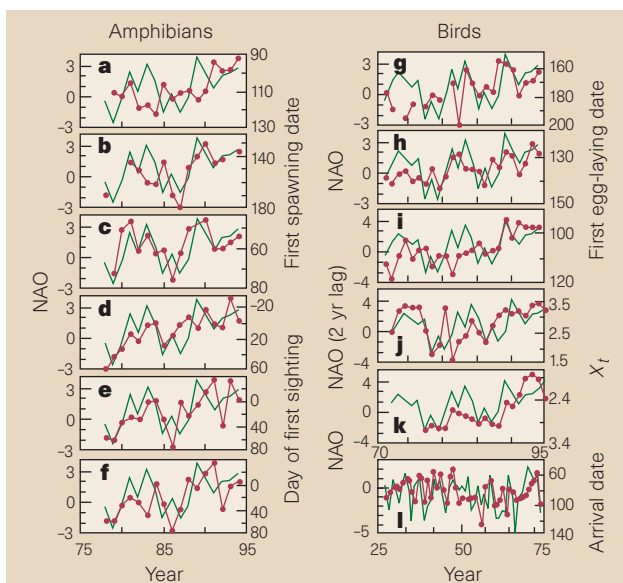
European amphibians and birds have been breeding consistently earlier over the past two to three decades^{1,2}. These changes have been attributed to the observed trends in increasing average spring temperatures in Europe³ producing earlier growing seasons⁴ and increased forage availability. Here we show that variations in breeding of Euro-

pean amphibians and birds are influenced by variations in a natural, large-scale atmospheric phenomenon, the North Atlantic Oscillation (NAO). Our results support the proximate cause (that is, increase in spring temperatures) of the altered breeding phenology as suggested previously^{1,2}, but by extending previous analyses as well as integrating data from other bird species, they also suggest that organisms with complex life histories respond to year-to-year variations in their abiotic environment.

The NAO determines most interannual

fluctuation in winter temperatures in the northern Atlantic region, and is correlated with global as well as regional temperature fluctuations^{5,6}. Also, the NAO influences the atmospheric export of dust from North Africa, thereby inducing significant fluctuations in regional radiative budgets in the North Atlantic region⁷. The increase in the NAO index over the past 20–30 years (Fig. 1) has contributed significantly to the increase in average winter temperature across Europe⁶. The contribution of the NAO to temperature changes complicates

Figure 1 Variations in amphibian and bird breeding. **a–f**, Temporal variation in first spawning date in *Bufo calamita* (**a**), *Rana kl. esculenta* (**b**) and *R. temporaria* (**c**), and temporal variation in first day of sighting in *Triturus vulgaris* (**d**), *T. helveticus* (**e**) and *T. cristatus* (**f**). **g–i**, Temporal variation in first egg-laying date in *Miliaria calandra* (**g**), *Phylloscopus collybita* (**h**) and *Pica pica* (**i**). **j–l**, Temporal variation in log-transformed breeding numbers (X_t) in UK populations of *Pluvialis apricari* (**j**) and *Actitis hypoleucos* (**k**), and in arrival date of a Norwegian population of *Alauda arvensis* (**l**). Superimposed on each figure is the NAO index (grey line). In coastal Europe, high, positive values of the NAO indicate warm, moist winters, whereas unusually cold, dry winters are indicated by low, negative NAO values⁹. Least-squares correlation coefficients between the NAO and breeding characteristics are, for the non-detrended and detrended (in brackets) time series, as follows (italic type indicates significance, $P < 0.05$): **a**, -0.13 (0.23); **b**, -0.52 (-0.44); **c**, -0.57 (-0.49); **d**, -0.59 (-0.33); **e**, -0.50 (-0.25); **f**, -0.41 (-0.39); **g**, -0.50 (-0.37); **h**, -0.56 (-0.50); **i**, -0.34 (-0.17); **j**, 0.41 (0.37); **k**, -0.68 (-0.51); **l**, -0.59 (-0.60). Detrending was performed by linear regression. All correlations were checked for autocorrelation. The NAO index was obtained from the Climate Analysis Section (Colorado, USA) internet home page (<http://www.cgd.ucar.edu:80/cas/climind/>). Data on amphibians are from ref. 1; data on first egg-laying dates are from ref. 2; data on UK breeding numbers are from refs 10, 11 and D.W. Yalden (unpublished); and data on arrival dates are from ref. 12.



the interpretation of climatic responses to greenhouse gases, so the possible links between atmospheric circulation and greenhouse gases need further exploration⁶.

Our analyses reveal that, after a period of increasingly warm winters (high positive NAOs from 1970 to 1994), both amphibians and birds bred consistently earlier, as documented in previous analyses^{1,2}. However, variations in breeding phenology of several species also showed significant year-to-year responses to fluctuations in the NAO about the trend (Fig. 1). This was particularly obvious for the skylark, where the data covered a 50-year period (Fig. 1l). Moreover, the NAO influenced annual breeding numbers in populations of golden plover (Fig. 1j) and common sandpiper (Fig. 1k).

By applying autoregressive analysis on breeding abundance with the NAO as a covariate (see ref. 8 for technical details), we found the most parsimonious autoregressive structures to be (coefficients $\pm$ s.e.m.) for plover:

$$X_t = (2.43 \pm 0.15) + (0.32 \pm 0.24)X_{t-1} - (0.39 \pm 0.24)X_{t-2} + (0.09 \pm 0.05)NAO_{t-1} + (0.14 \pm 0.06)NAO_{t-2} + (0.05 \pm 0.06)NAO_{t-3} + \epsilon_t$$

and for sandpiper:

$$X_t = (2.61 \pm 0.35) + (0.90 \pm 0.14)X_{t-1} - (0.06 \pm 0.02)NAO_{t-1} + \epsilon_t$$

where ϵ_t is a random variable with zero

mean and constant variance σ^2 . This analysis supports our correlations (Fig. 1j, k), but also demonstrates the simultaneous effect of density dependence, which, for the plover, counteracted the effect of the NAO. Whereas for plovers, cold winters (negative NAO) limit breeding¹⁰, for sandpipers warm, snowy winters (positive NAO) are limiting¹¹; the opposite effect of NAO on the two waders is confirmed by our analyses.

Our results, combined with recent studies on marine species showing the effect of NAO on pelagic primary production and zooplankton abundance in the North Sea⁹, thus add to the growing body of evidence^{13,14} on the traceable influence of climatic fluctuations on ecological processes.

Mads C. Forchhammer, Eric Post,

Nils Chr. Stenseth

Division of Zoology, Department of Biology,

University of Oslo, P.O. Box 1050 Blindern,

N-0316 Oslo, Norway

e-mail: n.c.stenseth@bio.uio.no

1. Beebe, T. J. C. *Nature* **374**, 219–220 (1995).
2. Crick, H. Q. P., Dudley, C., Glue, D. E. & Thomson, D. L. *Nature* **388**, 526 (1997).
3. Jones, P. D. *J. Climatol.* **7**, 1794–1802 (1994).
4. Myneni, T. B., Keelin, C. D., Tucker, C. J., Asrar, G. & Nemani, R. R. *Nature* **386**, 698–702 (1997).
5. Schlesinger, M. E. & Ramankutty, N. *Nature* **367**, 723–726 (1994).
6. Hurrell, J. W. *Science* **269**, 676–679 (1995).
7. Moulin, C., Lambert, C. E., Dulac, F. & Dayan, U. *Nature* **387**, 691–694 (1997).
8. Royama, T. *Analytic Population Dynamics* (Chapman & Hall, London, 1992).

9. Fromentin, J.-M. & Planque, B. *Mar. Ecol. Prog. Ser.* **134**, 111–118 (1996).
10. Yalden, D. W. & Pearce-Higgins, J. W. *Bird Study* **44**, 227–234 (1997).
11. Holland, P. K. & Yalden, D. W. *Bird Study* **38**, 151–159 (1991).
12. Lauscher, A. & Lauscher, F. *Phänologie Norwegens* vol. 4 (Eigenverlag, F. Lauscher, Vienna, 1990).
13. Jaksic, F. M., Feinsinger, P. & Jimenez, J. E. *Conserv. Biol.* **10**, 252–262 (1994).
14. Jaksic, F. M., Silva, S. I., Meserve, P. L. & Gutierrez, J. R. *Oikos* **78**, 341–354 (1997).

...temperature and egg-laying trends

Crick *et al.*¹ showed that the laying dates for many British birds have become earlier during 1971–95, which they ascribe to warmer springs resulting from global warming. We provide partial support for this view from a great tit population in Oxfordshire.

Since 1947, the Edward Grey Institute of Field Ornithology has maintained records of the breeding biology of great tits, *Parus major*, in Marley Wood, part of Wytham Wood near Oxford. Among the data collected are the annual laying dates (the mean date of the laying of the first egg by each pair) of the population. We also have access to detailed weather data collected by the University's Radcliffe Meteorological Station (courtesy of the School of Geography), about 3.5 km east of the centre of the study plot. Here we use, as a measure of spring temperature, the sum of the maximum temperatures for each day from 1 March to 25 April, which we have previously shown to be the best of several such measures².

Our data for the great tit conform with the finding of Crick *et al.*¹, for other species, that the laying date has become earlier since 1970 (Fig. 1a). However, we can find no evidence of such a trend before 1970; in the 24 years 1947–70 there was no significant trend towards earlier breeding, whereas in the 27 years since 1970 the trend towards earlier breeding is highly significant. Crick *et al.*¹ make the assumption that the earlier laying dates of the birds that they studied were due to increasing spring temperatures. We can confirm this for the great tit. Laying date is strongly related to spring temperatures (Fig. 1b), being earlier in years with high than with low temperature sums.

There is no evidence that the birds are now breeding earlier, relative to the spring weather, than they were formerly; the slope of the relationship has not changed between the two periods (Fig. 1b). The trend towards earlier breeding seems to come about solely because of the increasing temperatures in the spring (Fig. 1c). There seems to have been a large amount of variation in spring temperatures in the 1950s and up to 1962. In 1963–70 springs were rather cool, but since 1970 there has been a generally upward trend, accelerating

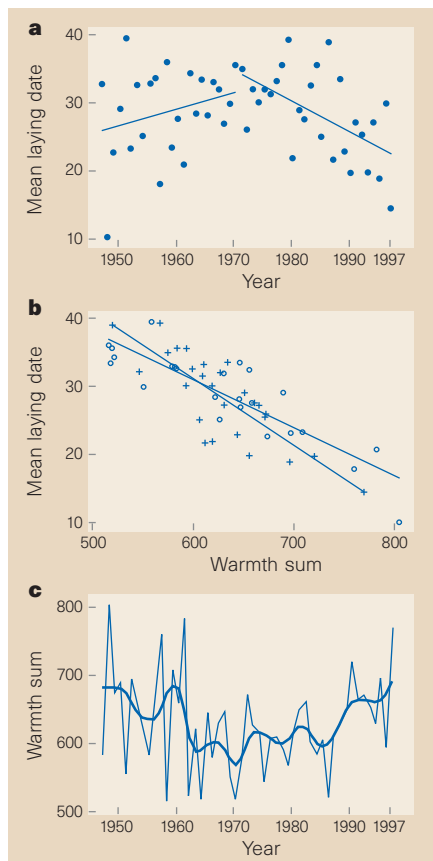


Figure 1 Laying date of Marley great tits and spring temperatures. **a**, Mean laying date (1 represents 1 April) plotted against time. Lines are regressions fitted to the two groups of data. The curvature of the best-fit model was confirmed by polynomial regression, which showed the quadratic term in time to be highly significant ($F=11.37$, d.f. = 1,47, $P=0.002$). The upward slope in 1947–70 is not significantly different from 0 (slope = 0.2035, s.e.m. = 0.2054, $t=0.99$, 21 d.f., $P=0.333$; if the exceptionally hot and early 1948 is removed the slope becomes 0.01). The tendency for laying to become earlier in 1971–97 is highly significant (slope = -0.4387 , s.e.m. = 0.1331, $t=3.30$, 25 d.f., $P=0.003$). **b**, Regression of mean laying date for great tits in Marley wood against warmth sum. The data are classified as 1947–70 and 1971–97. The warmth sum is calculated as the sum of the daily maxima from 1 March to 25 April in each year. The difference in slopes for 1947–70, -0.069 , and 1971–97, -0.097 , is not significant ($F=3.06$, d.f. = 1,47, $P=0.087$). The overall slope of -0.083 ± 0.008 s.e.m. is significantly different from 0 ($t=-10.42$, $P<0.001$). **c**, Temperature trends at Oxford. The thick line is the resistant smoother “4253H, twice”.

markedly from about 1985 onwards.

We know that there is considerable advantage to the birds to breed as early as they can, because the earliest breeders tend to produce the most surviving offspring³. However, the high energetic demands of breeding⁴ might constrain the time at which the females can start to lay if sufficient food is not available. Experiments involving the provision of artificial food in the spring^{5,6},

causing the birds in the experimental area to breed earlier than controls, support the idea that the timing of breeding is at least to some extent food-limited. Presumably, the early warm weather enables the birds to breed earlier because the food supply they need for breeding becomes available earlier.

R. H. McCleery, C. M. Perrins

Edward Grey Institute of Field Ornithology,
Department of Zoology,
University of Oxford,
South Parks Road, Oxford OX1 3PS, UK
e-mail: chris.perrins@zoo.ox.ac.uk

1. Crick, H. Q. P., Dudley, C., Glue, D. E. & Thomson, D. L. *Nature* **388**, 526 (1997).
2. Perrins, C. M. & McCleery, R. H. *Wilson Bull.* **101**, 236–253 (1989).
3. Perrins, C. M. *Ibis* **112**, 242–255 (1970).
4. Perrins, C. M. in *Proc. XXI Int. Ornithol. Congr.*, Vienna, August 1994, 1–14 (1996).
5. Nager, R. G., Rieger, C. & van Noorwijk, A. J. *Am. Ecol.* **66**, 495–507 (1997).
6. Brönnsen, A. von & Jansson, C. *Ornis Scandinavica* **11**, 173–178.
7. Velleman, P. F. & Hoaglin, D. C. *Applications, Basics, and Computing of Exploratory Data Analysis* (Duxbury, Boston, MA, 1981).

Synchronized courtship in fiddler crabs

The apparent paradox posed by the synchronization of mating displays by males competing to attract females has provoked considerable interest among evolutionary biologists^{1–3}. Such synchronized sexual signalling has only been documented for communication using light flashes (bioluminescence) or sound. It has been suggested that the “fundamental reasons that might favour precise adjustments in signal timing relative to that of a particular neighbour could only be compelling for signallers using these two channels”¹. Here we provide the first quantitative evidence for synchronous production of a conventional visual courtship signal, the movement of a body part.

The fiddler crab (*Uca annulipes*) lives in mixed-sex colonies on inter-tidal mudflats. When their habitat is exposed at low tide, crabs emerge from their burrows, feed and interact socially. Gravid females leave their own burrows and mate in the burrows of males⁴. Females find these burrows by sequentially entering (‘visiting’) the burrows of several males. Males attract females to their burrows by repeatedly raising and lowering their single greatly enlarged claw (97 ± 6% of ‘waving’ is directed at females, $n=30$ males). Several waving males usually cluster around a female. Before she mates, she visits one male from each of several clusters.

Preliminary field observations suggested that males in clusters wave synchronously⁵. To test this, we made video recordings of

females and clustered waving males and documented the timing (0.04 s precision) of waves by each male in the cluster relative to the visited male ($n=45$ clusters, each with 2–6 males). We define a wave cycle as the interval between the onset of successive claw waves of the visited male (wave cycle duration = T_v). Each wave of each of the other males in the cluster (‘neighbours’) was assigned to the wave cycle of the visited male in which it began. We calculated the difference in the onset time of the first wave of the visited male (t_v) and that of a neighbour (t_n), and the phase angle, $\alpha = ((t_n - t_v)/T_v) \times 360^\circ$ (Fig. 1a).

The phase angle is a measure of wave synchrony. If $\alpha = 0^\circ$ or 360° there is perfect synchrony, at $\alpha = 180^\circ$ waves are produced alternately. The mean T_v was 1.70 s (s.d. = 0.67); whereas the mean ($t_n - t_v$) was only 0.20 s (s.d. = 0.31; for both means, $n=328$ wave cycles by 45 visited males). Using circular statistics, we found the mean α ($\pm$ s.e.m.) for 328 pairs of waves was $4.4 \pm 2.9^\circ$. We then calculated the mean α for each of the 45 clusters and tested whether these were uniformly distributed, as expected if males in clusters wave independently of one another. The distribution was not uniform. Instead, these means were significantly concentrated around phase angles indicating close synchrony (Rayleigh test, $P<0.01$; mean $\pm$ s.e.m. per cluster = $6.8 \pm 18.7^\circ$). Moreover, in 23 of the 29 clusters where sample sizes were large enough to permit statistical testing, the phase angles were significantly clumped ($P<0.05$; Fig. 1b). Therefore, most neighbours, in most clusters, waved in close synchrony with the visited male.

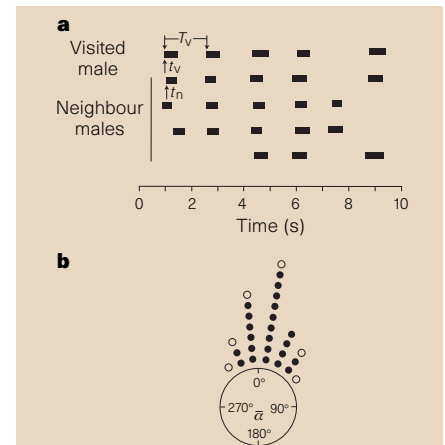


Figure 1 Synchronized courtship waving in clusters of male *Uca annulipes*. **a**, Waves (black bars) by five males in one of 45 clusters. The phase angle, α , is the difference in onset time of a wave of a neighbour male (t_n) and a wave of the male the female visited (t_v) as an angular proportion of the interval between onsets of the visited male's waves (T_v). **b**, Mean phase angles for 29 clusters. Filled circles: clusters with significantly non-uniform distributions; open circles: non-significant ($P>0.05$). Fieldwork was carried out in Durban Harbour, South Africa.

None of the proposed cooperative explanations for synchrony¹ seem to apply to waving in *U. annulipes*. First, there are no predators who would be confused by group synchrony thereby reducing each male's predation risk. Second, synchrony does not increase group conspicuousness to distant females because males wave synchronously only when a female is < 10 cm away (P. B., unpublished data). The most plausible explanation is one recently proposed for synchronous chorusing by a katydid². Modelling shows that competition between males to call before their neighbours can lead to synchrony. Males compete to call first because females prefer leading calls.

This so-called 'precedence effect', whereby signal receivers show greater responsiveness to the earlier signal in a pair, is found in many acoustic situations, including sound localization in humans. It has not, however, been reported in a visual communication system. In *U. annulipes*, the visited male produced leading waves significantly more often than his neighbours (4.5 ± 3.5 compared with 3.2 ± 3.3 waves; Wilcoxon test, $n = 45$, $P < 0.02$). Female *U. annulipes* may prefer leading signals and synchrony may arise as an incidental effect of competition between males to signal first.

This is the first example of synchronous production of a non-bioluminescent visual signal. Mechanisms of visual signal perception and processing must possess the properties that were previously thought to limit synchronized courtship signalling to acoustic and bioluminescent channels¹.

Patricia Backwell, Michael Jennions
Neville Passmore

Department of Zoology,
University of the Witwatersrand,
Braamfontein, Johannesburg 2050, South Africa
e-mail: backwelp@naos.si.edu

John Christy

Smithsonian Tropical Research Institute,
Apartado 2072, Balboa, Republic of Panama

- Greenfield, M. D. *Annu. Rev. Ecol. Syst.* **25**, 97–126 (1994).
- Greenfield, M. D. & Roizen, I. *Nature* **364**, 618–620 (1993).
- Sismondo, E. *Science* **249**, 55–58 (1990).
- Backwell, P. R. Y. & Passmore, N. I. *Behav. Ecol. Sociobiol.* **38**, 407–416 (1996).
- Gordon, H. R. S. *Am. Behav.* **134**, 238–241 (1958).

Connexin mutations and hearing loss

Kelsell *et al.*¹ provide convincing evidence that mutations in the gene encoding the gap-junction protein connexin 26 (Cx26) are responsible for autosomal recessive non-syndromic hearing loss at the DFNB1 locus on chromosome 13q12. They also report a small family with apparent autosomal dominant congenital hearing loss and autosomal dominant palmoplantar kerato-

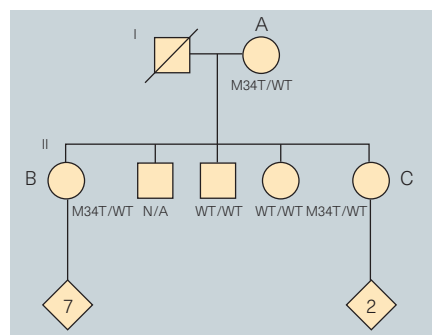


Figure 1 Pedigree structure in a family with the M34T variant of Cx26. Individual A (I:2) is 83 years old, individual B (II:1) is 57 years old and individual C (II:5) is 55 years old. The numbers in diamonds represent numbers of children. WT, wild-type; N/A, not available.

derma (PPK) in which two siblings with profound hearing loss are heterozygous for a single base-pair substitution resulting in a methionine-to-threonine (ATG-to-ACG) change in codon 34 (M34T) of Cx26 (refs 1, 2). The authors conclude that the M34T change is the genetic basis of profound hearing loss in this family and suggest, on this basis, that the Cx26 gene is responsible for autosomal dominant non-syndromic hearing loss (ADNSHL) at the DFNA3 locus chromosome 13q12 (ref. 3). We have identified a family in which the M34T variant is not associated with hearing loss, suggesting that this conclusion might be premature.

In a Cx26 mutation screen of 100 random individuals from the midwestern United States, we discovered one person who carries the M34T allele. DNA sequencing confirmed this person and two other family members to be heterozygous for this mutation. All three individuals have excellent hearing (Figs 1, 2)⁴. Individuals B and C report that their children, the youngest of whom is 24, all have normal hearing.

There are various explanations for these data. First, the M34T variant might be responsible for a form of ADNSHL that is not expressed in certain individuals. But the identification of multiple heterozygous individuals with no evidence of inherited

hearing loss argues against this suggestion. Second, individuals in this family might carry a compensatory change that nullifies the effects of the M34T variant. However, single-stranded conformational polymorphism (SSCP) and sequence analysis of the Cx26 coding region of individual C show no evidence of additional sequence variants. The most likely explanation for these data is that the M34T variant represents a simple polymorphism that does not cause autosomal dominant hearing loss and is present in a small percentage of the general population.

Determining whether a particular DNA variant represents a disease-causing mutation or a simple polymorphism requires information about the inheritance pattern of the DNA variant within affected families and the frequency of the variant allele within the general population. Although Kelsell *et al.* show that M34T segregates with the profound hearing-loss phenotype in one affected family, the few individuals available for study and the presence of a second deafness-associated disorder make it difficult to draw any clear conclusions based on this family alone^{2,5,6}. Kelsell *et al.* did not see the M34T variant in their screen of 80 chromosomes from non-related individuals, but a wider search might have revealed it.

Our results suggest that the M34T variant does not cause autosomal dominant hearing loss and emphasize the need for caution when interpreting mutation data based on a single affected family.

Daryl A. Scott, Michelle L. Kraft
Edwin M. Stone, Val C. Sheffield
Richard J.H. Smith

University of Iowa Hospitals and Clinics,
200 Hawkins Drive, Iowa City,
Iowa 52242-1078, USA
e-mail: richard-smith@uiowa.edu

- Kelsell, D. P. *et al.* *Nature* **387**, 80–83 (1997).
- Fitzgerald, D. A. & Verbov, J. L. *Br. J. Dermatol.* **134**, 939–942 (1996).
- Chaib, H. *et al.* *Hum. Mol. Genet.* **3**, 2219–2222 (1994).
- Gates, G. A., Cooper, J. C. Jr, Kannel, W. B. & Miller, N. J. *Behav. Techn. Audiol. Otol.* **11**, 247–256 (1990).
- Bititci, O. O. *J. Laryngol. Otol.* **89**, 1143–1146 (1975).
- Hatamochi, A., Nakagawa, S., Ueki, H., Miyoshi, K. & Iuchi, I. *Arch. Dermatol.* **118**, 605–607 (1982).

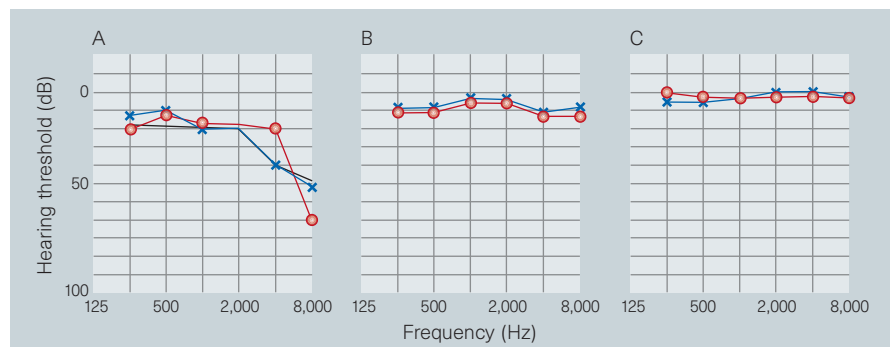


Figure 2 Pure-tone thresholds in the individuals identified in Fig. 1. Circles, right-ear thresholds; crosses, left-ear thresholds. The broken line in A represents the mean pure-tone thresholds in the better ear of women aged 80–84 (ref. 4).

Time and its milestones

Questioning the Millennium: A Rationalist's Guide to a Precisely Arbitrary Countdown

by Stephen Jay Gould
Harmony/Cape: 1997. Pp. 192.
\$17.95, £12.99

Calendrical Calculations

by Nachum Dershowitz and Edward M. Reingold
Cambridge University Press: 1997. Pp.307.
£40, \$64.95 (hbk); £14.95, \$22.95 (pbk)

E. G. Richards

With only two years to go to the next millennium, Stephen Jay Gould has been moved to ponder the what, when and why of this supposed watershed in human affairs. Gould points out that apocalyptic prophecies of the end of the world have been with us since the beginning of Christianity, and the idea of the Second Coming to be followed by the blissful years of the new millennium is part of the Christian tradition. Gould discusses a few of these ideas, which have appealed more to the poor and oppressed who welcome a change than to the Princes of the Church, happily ensconced in their palaces.

The Book of Revelation in the Bible suggests that history unfolds in 1,000-year cycles, or millennia. Several seventeenth-century divines addressed the date of the Creation, when the cycles started. A favourite approach was to tot up the ages of the biblical patriarchs and try to assign historical dates to biblical events.

Bishop Ussher made one of the best-known attempts. He calculated that the Creation began 4,000 years before the birth of Christ, at noon on 23 October 4004 BC. The odd four years are attributed to an error made by Dionysius Exiguus, who invented the Christian era in the sixth century. Thus Gould calculates that the Second Coming occurred a few weeks ago, on 23 October 1997. This date seemed to pass uneventfully, but the desire to celebrate the start of the two-thousandth year of the era remains, with a subtle change in the meaning of 'millennium'.

But when, exactly, should we open the champagne? Turning to calendrics, Gould points out that since about 1690 there has been an argument between those who supposed that the dawn of a new century should be celebrated on 1 January 1701 or on 1 January 1700, and so on. The proponents of high culture have, with few exceptions, espoused the former date, which marks the elapse of a whole number of centuries since the start of the Christian era, but the latter date, a nice round number, has been favoured by the populace. This is a sterile debate, and Gould wisely does not take sides.



Marking the passage of time: a Breton portolan and calendar showing the cycle of the Moon, 1546 (above) and a modern copy of an Aztec calendar (right).

Gould writes with humanity and with his usual flair and style. He provides a dozen or so illustrations of the Day of Judgement, and an index but no bibliography. He ends with a moving account of some idiot savants who, with impaired mental abilities, can nevertheless instantly calculate the day of the week of any date. For those convinced that our mental abilities can be encapsulated in a single parameter, this story provides a counterexample of someone who can perform a mental feat that eludes most but who is unable to master other mental skills. To find out who, you must read the book.

The French Revolution sparked off a widespread interest in the calendar and its reform. Many articles were published on the subject, including a method by Lewis Carroll (*Nature* 35, 517; 1887) for calculating the day of the week. The most recent comprehensive treatise on the calendars of the world, Friedrich Ginzel's *Handbuch der Mathematischen und Technischen Chronologie*, dates from 1914.

Since then we have seen the development of the digital computer, which now allows the instant transformation of a date in one calendar to a date in another. Nachum Dershowitz and Edward M. Reingold have provided us with a comprehensive set of computer programs to do just that, and, as a bonus, to calculate the dates of various festivals including Easter. They have chosen to present their algorithms in LISP, and one can download them from their World Wide Web page for one's personal use. But perhaps LISP



is not the most widely understood programming language, and some will regret that the algorithms are not presented in a more accessible form.

Dershowitz and Reingold consider a representative selection of arithmetic calendars in which a date, most often in the form of a triplet of numbers representing the day, month and year, is determined by well-defined arithmetical rules; the calendars considered include the Julian, Gregorian, Coptic, Ethiopian, Islamic, modern Persian, Baha'i, Jewish, Mayan and Old Hindu. They go on to a further set, the French Revolutionary, Chinese and Modern Hindu calendars, in which the start of a year or month is fixed by astronomical observation or at least by an accurate astronomical theory, and provide routines for translating each date into a day number and back again into a date. As the epoch of their day count, they chose the proleptic start of the Gregorian calendar, 1 January 1 AD. In this they differ from others

who have used the Julian day number, as used by astronomers.

Dershowitz and Reingold discuss briefly the history of each calendar, and provide selected references for further reading, and there are apposite illustrations. The logic of the calendars and the LISP algorithms are discussed adequately, if sometimes tersely. Some 100 pages of appendices are devoted to a full LISP listing and a table of selected days rendered in each of the calendars. They mention that there is at least one error, but point to an errata to be found on their Web page. Mayan calendricists should note that they have used the correlation of the epoch of the long count with Julian Day 584283, as opposed to 584285 which is favoured by some.

For those prepared to come to terms with the code, this book must surely become the standard work on calendar conversions. No historian, chronologist or recreational mathematician should be without it. □

E. G. Richards is at 2 Peckarmans Wood, London SE26 6RX, UK.

e-mail: egrichards@zetnet.co.uk

Centrepieces

Carl Sagan's Universe

edited by Yervant Terzian and Elizabeth Bilson

Cambridge University Press: 1997. Pp. 282.
£40, \$59.95 (hbk); £14.95, \$22.95 (pbk)

Emmanuel Lellouch

"The unifying theme of this symposium is the search for life elsewhere and the enhancement of life here. That captures the essence of Carl Sagan." Bruce Murray's words set the stage for *Carl Sagan's Universe*, the proceedings of a meeting organized in October 1994 to celebrate Sagan's achievements on the occasion of his sixtieth birthday (and two years before he died). The result is a collection of articles that reflect the unifying diversity of Sagan's interests.

The search for life in the Universe and for the origin of life on Earth is the backbone of the first half of the book. It provides a concise, often historical overview of the American and (former) Soviet space exploration programmes, and then a more general discussion of the conditions required for the development and maintenance of life.

Most of the articles are very well written and include good illustrations (particularly some early and rare pictures of Mars). Appropriate emphasis is given to the evolution of the environments of the Earth-like planets and to the role of comet and asteroid impacts, either for the delivery of biogenic elements or as planet-sterilizing events.

The importance of perspective and context in space exploration — one of Sagan's favourite themes — is stressed in sev-



Carl Sagan: tried to find life on other planets and to preserve the environment of our own.

eral chapters, bearing in mind its limitations. Would yet-to-be-discovered microfossils on Mars represent a separate martian origin of life, or would they instead suggest the inoculation of Mars by the Earth and/or the exchange of living organisms between the two worlds?

This first section reads essentially as a novel, and will be greatly enjoyed by both the planetary scientist — who will find in a concise, accurate and systematic form all the elements he already more or less knows — and the more general reader. The search for life then extends beyond the Solar System with a detailed historical and technical description of the search for extraterrestrial intelligence. There is then a (very difficult, not surprisingly) chapter on interstellar travel and time travel.

Science has shaped our world, not only through medical and technological revolutions, but also through its relationships with most human activities, including politics, power, education, communication and religion. The second part of *Carl Sagan's Universe* deals in depth with these interactions and illustrates the role of science as middleman in this multiplicity.

For centuries, religion and science have been, and still are to a large extent, unmerciful competitors. One of the most important achievements of science has been to challenge the concept of the centrality of mankind.

Sagan, in his own chapter, recalls that this idea of man being at the centre of the Universe and unique as a species was not only inherited from religious or philosophical teaching in Western civilizations, but is also a founding principle of most cultures. But this

'implicit' belief has been defied by science: astronomy has shown that the Sun is in "the obscure outskirts of an ordinary galaxy" and that other stars appear to have planets, and biology has shown that humans share some 99.6 per cent of their active genes with chimpanzees.

Joan Campbell has another way of looking at this, saying that "science offers facts for men and women of faith". Referring to the 1990 "Joint Appeal by Religion and Science for the Environment", she notes the common sense of awe and respect for the Universe shared by many scientists and those of faith, and suggests a partnership between science and religion aimed at preserving the Earth. This is all very well, but doesn't it look like an electoral alliance between two parties with remote positions?

Preserving the Earth was one of Sagan's great concerns. The book highlights his pioneering work on nuclear winter and its role in informing the public and decision-makers on the most important environmental questions.

On the issue of information, the talent of the speaker, the means of communication, and the general level of scientific education of the public are all crucial to success. Jon Lomberg's excellent chapter on the importance of the visual presentation of science reminded me of a public talk given by Sagan in 1991 outside the Division of Planetary Sciences meeting in Palo Alto, California. Sagan was describing the annual increase of the ozone hole over the South Pole. The region depleted in ozone was seen as an ugly, well-chosen brownish spot whose size varied with the seasonal and long-term variability of ozone. Each time the brown region

reached a maximum a distressed shudder went through the audience.

The book also emphasizes the interactions of science, politics and power which, at the end of a century dominated by military and space confrontation and the arms race, are obviously very strong. But the authors express different opinions on the plausibility of eliminating the nuclear danger from the world. The relationships between science and art are also briefly touched upon, but not too convincingly. Is it really science that influences art, or rather the object of science, that is, nature?

The general impression left by *Carl Sagan's Universe* is that, if mankind has not been given any centrality (only the merits of his actions, as Sagan says, may sometimes give him a justified feeling of importance), then science has certainly become central to mankind. Ann Druyan wonders whether democracy can exist without science.

The question raises the issue of whether, as science improves, it should be applied to encompass all aspects of human life. Sagan once proposed a position on abortion based on a purely scientific approach. His argument was logical and perfectly defensible, but is it the right approach to the problem? What about a scientific analysis of love, emotions, freedom or ethical behaviour? Regrettably, these issues are not addressed in the book.

Yet *Carl Sagan's Universe*, through the diversity of its contents and the generally high standard of its articles, remains a very fine piece of work that will interest a wide range of readers. □

Emmanuel Lellouch is at the Observatoire de Meudon, F-92195 Meudon, France.

How we feel

Friday's Footprint: How Society Shapes the Human Mind

by Leslie Brothers

Oxford University Press: 1997. Pp. 187.

£18.99, \$25

What Emotions Really Are: The Problem of Psychological Categories

by Paul E. Griffiths

University of Chicago Press: 1997. Pp. 286.

\$27.50, £21.95

Ray Dolan

The common theme of these two books is human emotion. *Friday's Footprint* explores the contribution of emotion to the shaping of a sense of self. *What Emotions Really Are* provides a dense exploration of our vernacular concepts of emotion. Both books, in their own ways, are important contributions to this expanding and important area of enquiry.

The title of *Friday's Footprint* turns out to

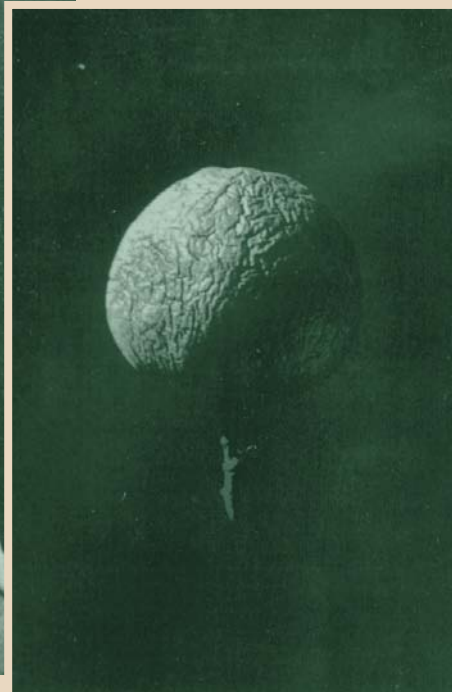
be an oblique reference to the fateful encounter in *Robinson Crusoe*. It echoes the author's longstanding interest and belief in the social basis of mental life. Leslie Brothers' conjecture that socialization is central to the development of a concept of self and others is bolstered by data from primate and human pathological studies. The latter include syndromes such as Capgras's, in which a person comes to believe that somebody close to them has been replaced by a physically identical double, and infantile autism, which is proposed to represent a failure to develop a concept of others.

In keeping with a social interactionist view of mind, Brothers proposes that brain circuitry associated with emotional and evaluative processing is central in the development of a sense of self. An important idea is that a dedicated neural circuit, referred to as an editor, is specialized for processing behaviourally relevant aspects of the social environment.

This editor is not some high-level central

system but a mechanism that signals the adaptive value of incoming sensory signals. The neural structure best fitted to meet the necessary requirements is the amygdala. Visual signals derived from the human face are put centre stage in view of its importance in emotional expression, paralinguistic display and verbalization.

What comes as a surprise is that Brothers has profound difficulties with the entire concept of emotion. Anathema to her is the implicit idea that emotion is an attribute of an isolated mind rather than, in her view, an expression of evolved mechanisms for communication and behavioural regulation. Much hostility is expressed towards the influential proposal for a fractionation of emotion into discrete subtypes as implicit in the concept of basic emotions. Brothers' ideas in this respect hark back to those of the behaviourist philosopher Gilbert Ryle, particularly in her formulation that what is distinct about emotions is their associated action tendencies. Just because speakers of



Scientific disciples of the artificial eye

One of the challenges of early photography was capturing the vast and intangible nature of astronomical objects. A novel approach was taken in a fascinating 1874 publication entitled *The Moon, Considered as a Planet, A World, and a Satellite* by the Scottish engineer James Nasmyth and the astronomer James Carpenter. Seeking to "educate the eye" about the "marvellous details" of the Moon, the authors based their book on drawings, astro-photographs and photographs of plaster

models. They made visual propositions about surface formations analogous to those of the Moon, using, for example, the pictures shown here of a wrinkled hand and apple.

This is but one of the many intriguing stories about the use of photography in science that are told in the richly illustrated *Beauty of Another Order* by Ann Thomas, with additional essays by M. Braun, M. Cazort, M. Kemp, J. P. McElhone and L. J. Schaaf (Yale University Press, \$50, £40).

English tactitly agree on the use of the word emotion, that does not guarantee its neurobiological validity.

In *What Emotions Really Are*, Paul Griffiths addresses the meaning of vernacular categories of emotion. Books on emotion often get bogged down in repetitive distillations of the multiple accounts of emotion that span the social to the biological. Griffiths's approach is refreshing in that he gets to the heart of the matter by dismissing as inadequate accounts of emotion subsumed by the general umbrella term of 'propositional attitudes'.

Griffiths also argues persuasively that cognitive theory explains very little about emotions. Emotional responses do not adjust themselves as readily as belief when new information about the environment is acquired. An example of this lack of fit between an emotional response and conscious evaluation would be a phobia. Griffiths echoes Brothers in identifying vernacular categories as a source of conceptual confusion and the absence of *a priori* reasons to assume that these categories have a necessary existence independent of language. A scientific psychology requires the refinement of emotion concepts to serve explanation and induction usefully.

Griffiths proposes that the general category of emotion should be replaced by, at a minimum, two distinct concepts. One would subsume the affect programs corresponding roughly to vernacular concepts of surprise, anger, fear, sadness, joy and disgust. The unity of this group is reflected in the fact that the associated affects are complex, coordinated and largely automated response patterns indexed in facial, musculoskeletal, vocal and autonomic nervous system outputs.

It should be stressed that affect programs do not explain all instances of the corresponding vernacular concepts. For instance, anger can vary from an affect program manifesting as spontaneous rage to more complex states as in plans for long-term redress of a wrong.

The latter type would best be subsumed under a second concept of emotion, distinct from the affect programs, termed the higher cognitive emotions, which account for emotions such as envy and shame. This second category describes what Griffiths terms irruptive motivations designed to enforce commitment to behavioural strategies that would not fit with the calculations of self-interest.

The higher cognitive emotions are also characterized by their ability to access beliefs and desires. In this regard, higher cognitive emotions are more flexible in expression than the highly stereotyped responses of affect programs.

Griffiths accepts that there is no adequate framework for research into higher cognitive

emotions but rejects any assumption that they represent an elaboration of more basic affects. His assumption is that, like the basic affects, these emotions are amenable to an evolutionary explanation.

It is difficult to do justice to Griffiths in a short review. His analysis of the concept of emotion and his proposal for the future direction of the field is the most compelling and best argued I have read. *What Emotions Really Are* makes a strong claim to be one of the best books to have emerged on the subject of human emotion. □

Ray Dolan is in the Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, UK.

Shadow history

To Light Such a Candle

by Keith J. Laidler

Oxford University Press: 1998. Pp. 384. £25

Peter Harman

In a series of historical accounts, drawn largely from developments in the physical sciences since the eighteenth century, Keith J. Laidler seeks to shed light on the relationship between science and technology. Despite the religious resonances of the title of this book, to which its epigraph draws attention, the 'candle' is that of the 'pure' science that has transformed material life during the past two centuries.

Laidler is concerned to study the past for its bearing on public attitudes and policy in the present. In illustration of his endeavour, he provides historical narratives of seven episodes in the history of science and technology: on James Watt, the steam engine and thermodynamics; on the history of photography; on Michael Faraday and electric power; on James Clerk Maxwell and radio transmission; on J. J. Thomson and the age of electronics; on the Braggs and molecular architecture, leading to the discovery of the structure of DNA and its implications; and on Albert Einstein, relativity and the quantum theory, leading to nuclear power and the atomic bomb.

Each chapter provides a historical narrative of a scientific quest, affirming the vitality of 'pure' science as an intellectual endeavour, and tracing the technological consequences of research, which were often not anticipated by the original protagonists.

Laidler states an intention to bridge the gulf between the perceptions of scientists and nonscientists about the very nature of the significant achievements of science during the past few centuries. He points out that, whereas scientists might point to Maxwell's field theory or quantum mechanics as examples of scientific advances, nonscientists would be apt to cite the invention of radio, colour television and the like.

Laidler's response to this discrepancy in perception, and to public ignorance, is couched in terms of a defence of research science as the 'candle' that has fired technology. He maintains that pure research should be judged on its own merits, that technology should be based on pure science and that "decisions about science and technology must be based on a careful consideration of all the factors involved".

These propositions are stated as 'conclusions', and although the first two seem bland and the third opaque, they are perfectly acceptable, if scarcely striking, as statements of opinion. The problem here is that, although Laidler's concerns and aims inspire sympathy, and his resolute defence of science as an intellectual quest can be amply documented, the argument of the book, except by broad inference, does little to establish his stated conclusions.

The substance of the book is given over, in Laidler's words, to "a 'popular' exposition of some aspects of science and technology". He proceeds by means of historical case studies, but his presentation is beset with difficulties. He relies on secondary sources, hardly surprising given the breadth of coverage in his book, but their use is not shaped by scholarly discrimination, and his text frequently meanders down seemingly irrelevant byways.

Moreover, the historical narrative is marred by errors and confusions, and although this is merely unfortunate in a popular study, it diminishes the authority of a work that aims to establish some broader thesis, as is the case here. But this thesis, on the nature and relationship of pure and applied science, is here only stated and addressed in general terms.

To Light Such a Candle therefore seems to fall between several stools. The historical analysis is not of a type to contribute effectively to a deeper understanding of the historical relationship of science and technology and their bearing on current public policy. Nor is this issue properly analysed. The reader is urged to conclude that, because science has had unexpected technological products, pure science should be encouraged as a way of securing future practical benefits. This may be reasonable, but it is scarcely original or striking, nor is it here rigorously argued.

The format of the book is attractive, and its argument provides a fairly accessible, but not always accurate or cogent, introductory exposition of topics in the history of science. But, given the fairly technical level of the text, it is likely to appeal to scientists rather than to the nonscientists whose confusion about the relationship between science and technology the author has sought to clarify. □

Peter Harman is in the Department of History, University of Lancaster, Lancaster LA1 4YG, UK.

Earthquakes and friction laws

Christopher H. Scholz

Earthquakes have long been recognized as resulting from a stick-slip frictional instability. The development of a full constitutive law for rock friction now shows that the gamut of earthquake phenomena—seismogenesis and seismic coupling, pre- and post-seismic phenomena, and the insensitivity of earthquakes to stress transients—all appear as manifestations of the richness of this friction law.

The traditional view of tectonics is that the lithosphere comprises a strong brittle layer overlying a weak ductile layer, which gives rise to two forms of deformation: brittle fracture, accompanied by earthquakes, in the upper layer, and aseismic ductile flow in the layer beneath. Although this view is not incorrect, it is imprecise, and in ways that can lead to serious misunderstandings. The term ductility, for example, can apply equally to two common rock deformation mechanisms: crystal plasticity, which occurs in rock above a critical temperature, and cataclastic flow, a type of granular deformation which can occur in poorly consolidated sediments. Although both exhibit ductility, these two deformation mechanisms have very different rheologies. Earthquakes, in turn, are associated with strength and brittleness—associations that are likewise sufficiently imprecise that, if taken much beyond the generality implied in the opening sentence, they can lead to serious misinterpretations about earthquake mechanics.

Lately, a newer, much more precise and predictive model for the earthquake mechanism has emerged, which has its roots in the observation that tectonic earthquakes seldom if ever occur by the sudden appearance and propagation of a new shear crack (or 'fault'). Instead, they occur by sudden slippage along a pre-existing fault or plate interface. They are therefore a frictional, rather than fracture, phenomenon, with brittle fracture playing a secondary role in the lengthening of faults¹ and frictional wear². This distinction was noted by several early workers³, but it was not until 1966 that Brace and Byerlee⁴ pointed out that earthquakes must be the result of a stick-slip frictional instability. Thus, the earthquake is the 'slip', and the 'stick' is the interseismic period of elastic strain accumulation. Subsequently, a complete constitutive law for rock friction has been developed based on laboratory studies. A surprising result is that a great many other aspects of earthquake phenomena also now seem to result from the nature of the friction on faults. The properties traditionally thought to control these processes—strength, brittleness and ductility—are subsumed within the overarching concept of frictional stability regimes.

Constitutive law of rock friction

In the standard model of stick-slip friction it is assumed that sliding begins when the ratio of shear to normal stress on the surface reaches a value μ_s , the static friction coefficient. Once sliding initiates, frictional resistance falls to a lower dynamic friction coefficient, μ_d , and this weakening of sliding resistance may, depending on the stiffness of the system, result in a dynamic instability. Following the suggestion of Brace and Byerlee, a great deal of attention was focused on the physics of rock friction, from which it was found that most of the statements in the standard model had to be revised. First, it was found that μ_s depends on the history of the sliding surface. If the surfaces are in static contact under load for time t , then μ_s increases slowly as $\log t$ (ref. 5). Second, the dynamic friction, when measured in the steady-state sliding regime, depends⁶ on the sliding velocity, V . This dependence, which goes as $\log V$, may be either positive or negative, depending on the rock type and certain other parameters such as temperature⁷. Last, if subjected to a sudden change in sliding velocity, friction is

found to evolve to its new steady-state value over a characteristic slip distance $\mathcal{L}$ (refs 8, 9).

The ageing of μ_s and the velocity dependence of μ_d are related behaviours¹⁰ which result from creep of the surface contact and a consequent increase in real contact area with time of contact^{11,12}. The critical slip distance $\mathcal{L}$ is interpreted as a memory distance over which the contact population changes^{9,13,14}. All of the experimental results are well described by an empirical, heuristic model known as the rate- and state-variable constitutive law, outlined in Box 1. This

Box 1 Rate- and state-variable friction law

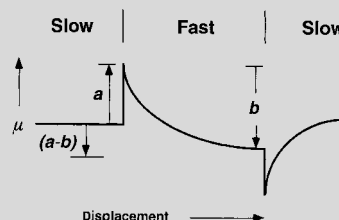
There are several forms of rate/state-variable constitutive law that have been used to model laboratory observations of rock friction. The version currently in best agreement with experimental data²⁰, known as the Dieterich-Ruina or 'slowness' law, is expressed as

$$\tau = \left[\mu_0 + a \ln \left(\frac{V}{V_0} \right) + b \ln \left(\frac{V_0 \theta}{\mathcal{L}} \right) \right] \bar{\sigma} \quad (1)$$

where τ is shear stress and $\bar{\sigma}$ is effective normal stress (applied normal stress minus pore pressure). In the bracketed friction term, V is slip velocity, V_0 a reference velocity, μ_0 the steady-state friction at $V = V_0$, and a and b are material properties. $\mathcal{L}$ is the critical slip distance and the state variable, θ , evolves according to:

$$\frac{d\theta}{dt} = 1 - \frac{\theta V}{\mathcal{L}} \quad (2)$$

The significance of these various terms is illustrated in the diagram below, which shows schematically but faithfully the experimentally observed frictional response to a suddenly imposed e -fold increase and then decrease in sliding velocity.



On initial application of the rate increase there is an increase a in friction, known as the direct velocity effect. This is followed by an evolutionary effect involving a decrease in friction, of magnitude b .

The friction at steady state is:

$$\tau = \left[\mu_0 + (a - b) \ln (V/V_0) \right] \bar{\sigma} \quad (3)$$

from which equation (1) in the main text arises. There is a continuum of changing friction values, but if dynamic friction, μ_d , is defined as steady-state friction at velocity V , then $d\mu_d/d(\ln V) = a - b$. Similarly, if static friction μ_s is defined as the starting friction following a period of time t in stationary contact, then for long t , $d\mu_s/d(\ln t) = b$. The name 'slowness law' arises because at steady state, the state variable is proportional to slowness, $\theta_{ss} = \mathcal{L}/V$. The critical slip distance $\mathcal{L}$ is often interpreted as the sliding distance required to renew the contact population. In this view θ_{ss} represents an average contact lifetime.

form of friction does not seem to be very material dependent: it also applies to some metals⁸ and to paper, wood and some plastics^{15,16}. The former distinction between μ_s and μ_d disappears in this model. The base friction μ_0 has a value nearly independent of rock type and temperature^{7,17}. It is modified by second-order effects involving a dependence on sliding velocity and a state variable θ , and it is these second-order effects that result in the interesting modes of behaviour discussed here. The base friction, which determines the frictional strength of the fault, does not concern us in this discussion. The fault strength is not involved in the seismogenic behaviour of the fault, which is solely determined by its frictional stability, not its strength. Fault strength does play a role in frictional heating of faults, which can produce several interesting effects^{10,18} that will not be discussed here.

Friction stability regimes and seismogenesis

Frictional stability depends on two friction parameters, $\mathcal{L}$ and the combined parameter $(a - b)$, defined as the velocity dependence of steady-state friction (a and b are defined in Box 1):

$$a - b = \frac{\partial \mu^{ss}}{\partial [\ln(V)]} \quad (1)$$

The frictional stability regimes are described in Box 2. If $(a - b) \geq 0$, the material is said to be velocity strengthening, and will always be stable. In the velocity-weakening field, $(a - b) < 0$, there is a Hopf bifurcation between an unstable regime and a conditionally stable one. Considering a simple spring–slider model with fixed stiffness k , the bifurcation occurs at a critical value of effective normal stress $\bar{\sigma}_c$, given by:

$$\bar{\sigma}_c = \frac{k\mathcal{L}}{-(a - b)} \quad (2)$$

If $\bar{\sigma} > \bar{\sigma}_c$, sliding is unstable under quasistatic loading. In the conditionally stable regime, $\bar{\sigma} < \bar{\sigma}_c$, sliding is stable under quasistatic loading but can become unstable under dynamic loading if subjected to a velocity jump exceeding ΔV , as shown in Box 2. In a narrow region at the bifurcation, sliding occurs by a self-sustaining oscillatory motion^{6,15,19} (shaded region, in the second diagram in Box 2). Although the friction law described in Box 1 can be written in several ways which differ in detail²⁰, those details do not influence the above definitions of the stability states, which control the seismic behaviour of faults discussed here.

The three stability regimes have the following consequences for earthquakes. Earthquakes can nucleate only in those regions of a fault that lie within the unstable regime. They may propagate indefinitely into conditionally stable regions, provided that their dynamic stresses continue to produce a large enough velocity jump. If earthquakes propagate into a stable region, on the other hand, a negative stress drop will occur, resulting in a large energy sink that will rapidly stop the propagation of the earthquake.

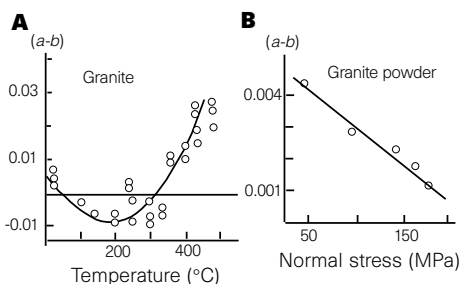


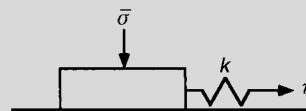
Figure 1 Systematics of the friction parameter $(a - b)$. **A**, Dependence of $(a - b)$ on temperature for granite (from refs 7, 21). **B**, Dependence of $(a - b)$ on pressure for granulated granite (ref. 24). This effect, due to lithification, should be augmented with temperature.

The primary parameter that determines stability, $(a - b)$, is a material property (see Box 1). The main systematics of this parameter that concern us are summarized in Fig. 1. Figure 1A shows the dependence of $(a - b)$ on temperature for granite^{7,21}. It is negative at low temperatures and becomes positive for temperatures above about 300 °C. This transition temperature corresponds to the onset of crystal plasticity of quartz, the most ductile of the major minerals in granite²². It may be a general statement that for low-porosity crystalline rocks a transition from negative to positive $(a - b)$ corresponds to a change from elastic-brittle deformation to crystal plasticity in the micro-mechanics of friction. As another example, halite (rock salt), a much more ductile mineral, undergoes the same two corresponding transitions at 25 °C and a pressure of about 70 MPa (ref. 23). These observations indicate that for faults in granite, the representative rock of the continental crust, we should not expect earthquakes to occur below a depth at which the temperature is 300 °C, and faults in salt should be aseismic under almost all conditions.

Faults are not simply frictional contacts of bare rock surfaces: they are usually lined with seal, detritus, called cataclastite or fault gouge. The shearing of such granular material involves an additional hardening mechanism (involving dilatancy), which tends to make $(a - b)$ more positive²⁴. For such materials, $(a - b)$ is positive when

Box 2 Stability regimes

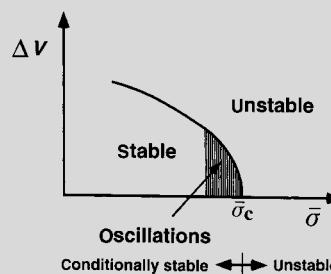
Consider a simple spring–slider model with spring stiffness k , as shown in the diagram below, in which the slider obeys the rate/state-variable friction law.



The stability of this system depends entirely on $\bar{\sigma}$, τ , k , the friction parameters $(a - b)$ and $\mathcal{L}$, and is independent of base friction μ_0 . The following are the conditions defining the stability regimes.

$(a - b) > 0$. This is velocity-strengthening behaviour, which is intrinsically **stable**. No earthquake can nucleate in this field, and any earthquake propagating into this field will produce there a negative stress drop, which will rapidly terminate propagation.

$(a - b) < 0$. The stability diagram for the system exhibiting velocity weakening is shown in the diagram below.



This diagram shows the velocity jump, ΔV , necessary to destabilize the system as a function of the applied normal stress $\bar{\sigma}$. If $\bar{\sigma} \geq \bar{\sigma}_c$, as defined by equation (2) in the main text, the system is unstable with respect to a vanishing velocity perturbation ΔV —that is, to quasistatic loading. This is the **unstable** field. If the effective normal stress is less than the critical value, it requires a finite velocity ‘kick’ to become unstable. Thus, in this **conditionally stable** field, the system is stable under quasistatic loading but may become unstable under sufficiently strong dynamic loading. Earthquakes may nucleate only in the unstable field, but may propagate into the conditionally stable field. At the border of the stability transition there is a narrow region in which self-sustaining oscillatory motion occurs, as indicated by the shaded region.

the material is poorly consolidated, but decreases at elevated pressure and temperature as the material becomes lithified (Fig. 1B). Therefore faults may also have a stable region near the surface, owing to the presence of such loosely consolidated material²⁵.

These considerations allow the construction of synoptic models for the two primary sites of tectonic earthquakes, crustal faults and subduction zone interfaces, as shown in Fig. 2. In the centre of this figure is drawn the expected variation of the friction stability parameter $\zeta = (a - b)\bar{\sigma}$. This is positive at shallow depths because of the presence of unconsolidated granular material and at large depths because of the onset of plasticity at a critical temperature—hence, regions above and below these stability transitions are stable (shown blue). The regions in which the stability parameter exceeds the threshold defined in equation (2) are unstable, indicated by red, and yellow indicates the regions of conditional stability. Thus, the red regions define the seismogenic zone, the depth range over which earthquakes may nucleate, as indicated by their hypocentral depths, an example of which is given on the right of Fig. 2.

For crustal faults, the upper transition depth is typically observed to be at 3–4 km, but may be absent at faults on which there has been little slip and hence little or no gouge developed²⁵. The lower transition occurs at 15–20 km, corresponding to the onset of plasticity of quartz at about 300 °C. The depth at which this occurs depends on the local thermal gradient²⁶. For subduction zones the upper transition occurs at the base of the accretionary prism of scraped-off oceanic sediments, where it encounters a ‘backstop’ of competent rock²⁷. Because the thickness of the sedimentary wedge is quite variable, so is the depth of this transition—it may be as deep as 10 km. The lower transition occurs at depths as great as 45 km at subduction zones. This greater depth is a result of lower thermal gradients, owing to the subduction of the cold oceanic plate, although this may vary widely because of variations in the age of the subducting plate, which strongly affects the thermal regime²⁸. The transition is also deeper because the basalt of the oceanic plate contains no quartz—the most ductile mineral in basalt is feldspar, which becomes plastic at about 450 °C (ref. 22). Because the seismogenic zone is much wider than for crustal faults—up to 150 km—and because they tend to be more continuous along strike, subduction zones produce by far the largest earthquakes in the world.

If a large earthquake occurs on a crustal fault, it will often have enough energy to propagate through the narrow shallow stable region and breach the surface. It may also often propagate a short distance into the ductile stable region at depth, for which there is geological evidence^{29,30}. However, for subduction zones with a wide accretionary prism, large earthquakes will often not breach the

surface. Whether they do or not is thought to be important in determining how efficient they are in generating tsunamis³¹.

Seismic coupling and seismic styles

The linear measure of earthquake size is seismic moment, $M_0 = GuA$, where u is the mean slip in the earthquake, A the rupture area and G the shear modulus. The moment release rate of a fault or plate boundary is thus $\dot{M}_0 = GvA$ where v is the long-term slip velocity and A is now the total fault area. We define the seismic coupling coefficient χ as the ratio of the moment release rate determined from summing earthquakes to the total rate obtained by determining v from a plate-tectonic model or geological data. The parameter χ is a good measure of the overall stability state of a fault. If the fault is entirely in the unstable field, $\chi = 1$, and if entirely in the stable field $\chi = 0$; otherwise, χ will be somewhere in between.

For most crustal faults, χ is indistinguishable from 1; that is, all of the fault slip occurs during earthquakes and these faults are said to be fully seismically coupled. An important exception is the so-called ‘creeping’ section of the San Andreas fault, a 170-km-long stretch in central California where the fault slips aseismically. Much of this aseismic slip occurs as ‘creep episodes’ (Fig. 3), which appear to be the same as the oscillatory behaviour observed at the stability boundary—*prima facie* evidence that this part of the fault is in the conditionally stable regime close to the bifurcation of equation (2). It is sufficiently far from the stability boundary, though, to prevent earthquakes on neighbouring sections of the fault from propagating very far into this region. The most likely mechanism for the anomalous behaviour of this section of the fault is the presence there of unusually high pore pressures in the fault zone³². I note that the effective normal stress is $\bar{\sigma} = (\sigma - p)$, where σ is the applied normal stress and p the pore pressure. If p approaches σ , the stability parameter ζ may be reduced so that the entire depth range of the fault normally in the unstable (red) field in Fig. 2 is shifted to the yellow field.

Although such seismic decoupling seems to be rare for crustal faults, it is not rare for subduction zones, which vary from being fully coupled to almost entirely decoupled³³. The difference seems to be due to the stress state. Stress measurements in deep boreholes³⁴ in continents typically show that deviatoric stresses increase with depth so that at all depths the stresses are just below that necessary to cause sliding on a favourably oriented fault with a friction coefficient consistent with laboratory values (about 0.6). The pore pressures observed in deep boreholes usually increase with the hydrostatic gradient, and the vertical stress usually increases with the weight of the overlying rock. The most-studied fault, the San Andreas of California, seems to be exceptional, in that it seems to be

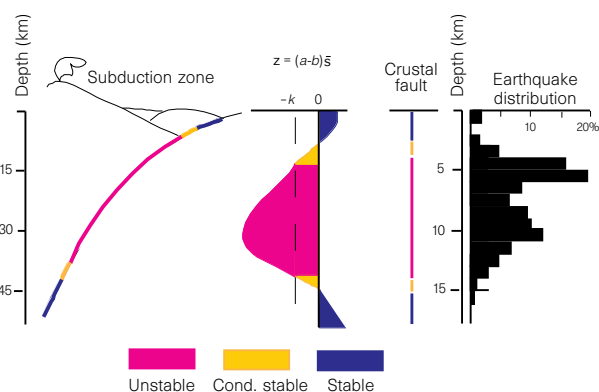


Figure 2 A synoptic model for stability as a function of depth for crustal faults and subduction zones. The central panel and the crustal fault model are taken from ref. 22; the subduction zone model (left) is from refs 27, 28; the histogram of the depth distribution of earthquakes (right) is for a section of the San Andreas fault near Parkfield, California (data from ref. 25).

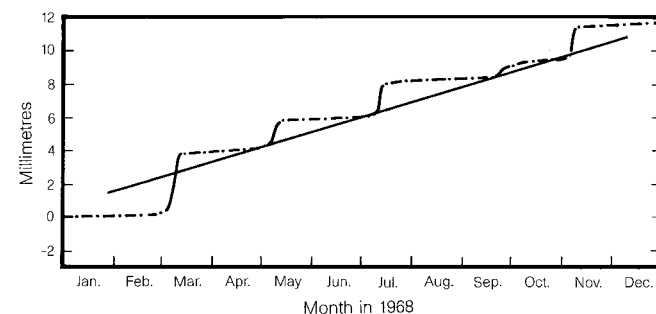


Figure 3 Oscillatory motion (creep episodes) of the creeping section of the San Andreas fault in central California (from ref. 10). The straight line is for reference.

sliding under very low shear stresses³⁵. This can be consistent with the friction law only if the pore pressure within the fault is near the lithostatic load (weight of overburden), which is very problematical³⁶. The important point is that crustal faults are subjected to remotely applied loads, and the effective normal stress on them is mainly determined by the lithostatic load minus the pore pressure, which is in the more usual case the hydrostatic head. For subduction zones, on the other hand, the forces that drive the plates are local to the subduction zone and may vary widely, which results in great variation of the effective normal stress supported by the plate interface. An analysis of the reduction of normal force (relative to a standard state) applied across subduction interfaces³⁷, calculated from the plate-tectonic driving forces, is shown in Fig. 4 for most of the world's subduction zones. The seismic coupling coefficient χ , determined from seismicity data, decreases from high to low values at a critical value corresponding to $\bar{\sigma}_c$, which was determined independently of the data shown in the figure. Owing to the shortness of the seismic record, these values of χ are not very well determined³⁸, but they are good enough to allow one to distinguish the coupled from the decoupled zones. Thus coupled and decoupled subduction zones are on either side of the stability transition boundary. On a local scale, irregularities caused, for example, by the subduction of seamounts can produce local increases in normal stress with the result that otherwise decoupled subduction zones may become locally coupled³⁹.

The three stability states result in three distinctive seismic styles. Regions with the stable field, such as the outer parts of accretionary prisms and faults in salt, are totally aseismic^{27,40}. Faults in the unstable field are characterized by infrequent large earthquakes separated by long interseismic periods of quiescence. On the other hand, faults in the conditionally stable regime, such as the creeping section of the San Andreas and the decoupled subduction zones, are characterized by high steady rates of small-event activity and no large events ('large' events are earthquakes that rupture the entire seismogenic thickness). These small events together contribute very little to the total moment release, which is primarily aseismic^{39,41}. Small events are found to occur repeatedly with a high repetition rate at the same spots⁴². These spots may mark small geometric irregularities⁴³ where the normal stress is higher, causing a transition to the unstable field.

Stages in the seismic cycle

As noted above, seismically coupled faults are typified by infrequent

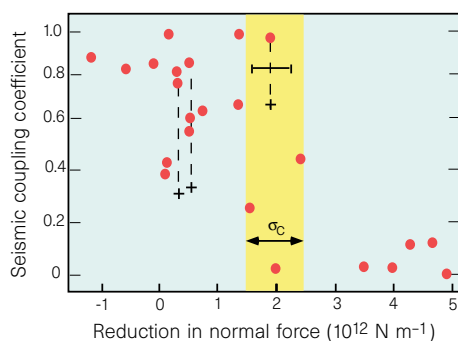


Figure 4 The observed seismic coupling coefficient χ versus the calculated reduction in normal force from a standard state for most of the Earth's subduction zones (after ref. 37). The transition point, $\bar{\sigma}_c$, was independently estimated from the Izu-Bonin/Mariana arc.

large events, separated by long quiescent interseismic periods in which the stresses relaxed by the preceding earthquake are restored. A frictional model of the seismic cycle of a strike-slip fault^{44,45}, purposely reminiscent of the San Andreas fault, is shown in Fig. 5. In this two-dimensional model, the fault is driven remotely at a constant velocity; the figure shows the slip on the fault as a function of depth at different times during the seismic cycle. The only assumption in the model is that it obeys the friction law of Box 1 with the $(a - b)$ parameter varying as in Fig. 1A and $\bar{\sigma}$ increasing with depth in a way consistent with the borehole data summarized above. The depth of the transition from unstable to stable regimes is at 11 km, in accordance with the geothermal gradient typical of the San Andreas fault.

During the interseismic period (shown blue in Fig. 5), the fault is loaded by steady slip on the deep, stable portion of the fault. Just before the earthquake a pre-seismic phase, known as nucleation, occurs (orange); in this phase, slip accelerates until the instability results in the coseismic motions (red). These penetrate below the stability boundary, reloading that region, which relaxes in a post-seismic phase of accelerated deep slip (green), at a rate that decays exponentially with time within a few years to a decade following the mainshock. Geodetic data strongly support the main features of this model⁴⁶⁻⁴⁸; the interseismic strain accumulation resulting from deep slip below a locking depth, above which the coseismic slip occurs, and a postseismic relaxation phase with decelerating deep slip⁴⁹. The pre-seismic nucleation phase is sometimes associated with the occurrence of foreshocks⁵⁰, and may also be responsible for certain other precursory phenomena that have been occasionally observed¹⁰.

A shallow relaxation phenomenon called afterslip is often observed, in which a fault slips aseismically at the surface in proportion to the logarithm of the time elapsed since an earthquake just below. Afterslip is usually observed where a thick layer of poorly consolidated sediments overlies the fault, which partially or totally impedes the earthquake from breaching the surface. Afterslip can be described by a model similar to that illustrated in Fig. 5 but with a thick stable layer at the top⁵¹; an example is shown in Fig. 6. This phenomenon has also been observed in partially coupled subduction zones, where an earthquake in an unstable patch drives afterslip in adjacent conditionally stable or stable regions⁵². The other typical postseismic phenomena, aftershock sequences, which obey a well defined hyperbolic decay law known as the Omori law, have also been shown to be a prediction of the rate- and state-variable friction

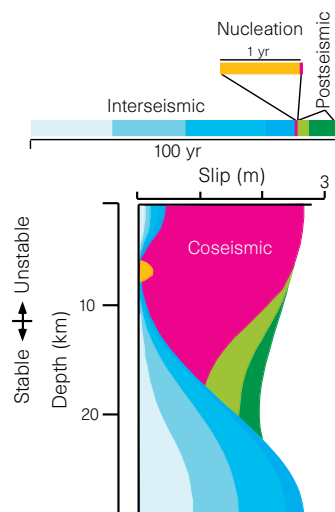


Figure 5 Slip as a function of depth over the seismic cycle of a strike-slip fault, using a frictional model containing a transition from unstable to stable friction at 11 km depth (after ref. 44).

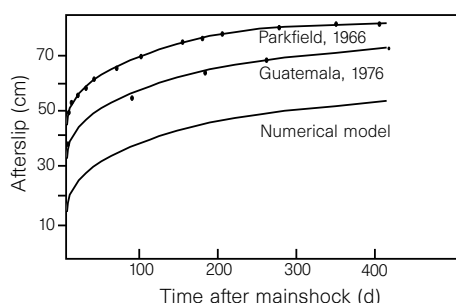


Figure 6 Afterslip observed at the surface following two large strike-slip earthquakes compared with the results of a numerical model of slip in a stable layer overlying an unstable region which slipped coseismically (from ref. 51).

law⁵³. This part of the theory has also been successfully tested with observations⁵⁴.

When the instability condition for a one-dimensional spring slider, equation (2), is generalized for the two- or three-dimensional case of a slipping patch of size L , the stiffness, k , scales inversely with L according to $k = \eta G/L$, where G is the shear modulus and η is a geometrical constant of the order of unity. This implies that the instability occurs when the slipping patch reaches a critical size L_c , the nucleation length, given by:

$$L_c = \frac{G\eta L}{(b-a)\bar{\sigma}} \quad (3)$$

Modelling⁵⁵ and laboratory observations^{6,56} of this nucleation process indicate that stable sliding initiates at a point and then spreads out with an accelerating sliding velocity until the instability arises at L_c . Whether or not such nucleation occurs on natural faults, and whether it is large enough to be detected, is central to the problem of short-term earthquake prediction. However, the physical significance and scaling of the key parameter L is not known. It is very small ($\sim 10 \mu\text{m}$) in the laboratory. Various attempts to model L , assuming that it is a property of the surface contact topography⁵⁷ or gouge zone thickness⁵⁸, suggest that it may be much larger at the fault scale. If L_c is a constant for natural faults, it also represents a minimum earthquake dimension. In that case $L_c < 10 \text{ m}$, the rupture size of the smallest observed earthquakes⁵⁹. On the other hand, observations of the dimensions of foreshock zones and of a precursory seismic phase both indicate that nucleation length may be of the order of kilometres and scale with the size of the ensuing earthquake^{60,61}.

Earthquake insensitivity to transients

There is one last property of this friction law which clears up some long-known mysteries about earthquakes. The direct friction effect (Box 1) and the finite size and duration of nucleation prohibit earthquakes from being triggered by high-frequency stress oscillations^{62,75}. Hence, as has been repeatedly verified over the past 75 years, earthquakes are not triggered by Earth tides⁶³. They are also not triggered (except in magmatic systems⁶⁴) by the seismic waves from other earthquakes⁶⁵, even though the smaller residual static stresses from earthquakes may trigger earthquakes following some time delay⁶⁶.

Outstanding problems in earthquake mechanics

There are, of course, many details yet to be worked out about the friction law, its parameters, and their scaling properties and applications to natural seismic phenomena. The scaling of L and its effect on nucleation is but one of these. But, lest the above discussion give the impression that most of what is worth knowing about earthquake mechanics is now well understood, I will describe several important problems that are currently at issue.

One important question is: what gives rise to the complexity of earthquakes? Earthquake populations obey a very well defined power-law size distribution, known as the Gutenberg–Richter law—such power laws being the hallmarks of systems exhibiting ‘self-organized criticality’ (ref. 76). The internal dynamics of earthquakes also exhibit complexity, with very broad-band velocity and acceleration spectra, while at the same time obeying well defined self-similar scaling laws in terms of the static parameters¹⁰. Is this complexity a result of the heterogeneity of faults, which have quasi-fractal scaling of surface topography⁶⁷, or of the nonlinearity of the friction laws, or both?

Dynamic models of arrays of spring-coupled block sliders^{68–70} obeying simple versions of these friction laws have been successful in reproducing Gutenberg–Richter statistics and many other aspects of observed earthquake complexity. These models assume no built-in heterogeneity, hence these results imply that this complexity may arise solely from the nonlinearity of the friction. On the other hand, continuum models show that complexity does not arise easily from the friction alone^{45,71}, unless extreme values for the friction parameters are assumed⁷². But these studies are far from exploring all regimes of friction parameter space. Furthermore, as remarked on below, the friction laws are still quite simple: there may be other aspects not yet uncovered in laboratory experiments.

There is a newly discovered class of earthquakes that are not expected from the friction laws as currently formulated: these are the ‘slow’ earthquakes, which are characterized by moment release rates much lower than those seen in other earthquakes⁷³. A slow earthquake occurring at a subduction zone may generate a much larger tsunami than expected from its moment as measured in the usual frequency band⁷⁴. One might suppose that some additional dissipative term, not yet recognized in the laboratory, may be present in order to explain this kind of earthquake.

Although there are many interesting phenomena yet to be explored, the success of the rate- and state-variable friction laws, simple as they are, in explaining a wide range of earthquake phenomena gives confidence that they will provide the basis for many exciting future discoveries as well. □

Christopher H. Scholz is at the Lamont-Doherty Earth Observatory and Department of Earth and Environmental Sciences, Columbia University, Palisades, New York 10964, USA

- Cowie, P. A. & Scholz, C. H. Growth of faults by the accumulation of seismic slip. *J. Geophys. Res.* **97**, 11085–11095 (1992).
- Scholz, C. H. Wear and gouge formation in brittle faulting. *Geology* **15**, 493–495 (1987).
- Gilbert, G. K. A theory of the earthquakes of the Great Basin, with a practical application. *Am. J. Sci.* **XXVII**, 49–54 (1884).
- Brace, W. F. & Byerlee, J. D. Stick slip as a mechanism for earthquakes. *Science* **153**, 990–992 (1966).
- Dieterich, J. Time-dependence of rock friction. *J. Geophys. Res.* **77**, 3690–3697 (1972).
- Scholz, C., Molnar, P. & Johnson, T. Detailed studies of frictional sliding of granite and implications for the earthquake mechanism. *J. Geophys. Res.* **77**, 6392–6406 (1972).
- Stesky, R. *et al.* Friction in faulted rock at high temperature and pressure. *Tectonophysics* **23**, 177–203 (1974).
- Rabinowicz, E. The intrinsic variables affecting the stick-slip process. *Proc. Phys. Soc. (London)* **71**, 668–675 (1958).
- Dieterich, J. Time dependent friction and the mechanics of stick slip. *Pure Appl. Geophys.* **116**, 790–806 (1978).
- Scholz, C. H. *The Mechanics of Earthquakes and Faulting* (Cambridge Univ. Press, 1990).
- Scholz, C. H. & Engelder, T. Role of asperity indentation and ploughing in rock friction. *Int. J. Rock Mech. Min. Sci.* **13**, 149–154 (1976).
- Wang, W. & Scholz, C. H. Micromechanics of the velocity and normal stress dependence of rock friction. *Pure Appl. Geophys.* **143**, 303–316 (1994).
- Dieterich, J. Modelling of rock friction: 1. Experimental results and constitutive equations. *J. Geophys. Res.* **84**, 2161–2168 (1979).
- Ruina, A. L. Slip instability and state variable friction laws. *J. Geophys. Res.* **88**, 10359–10370 (1983).
- Heslot, F., Baumberger, T., Perrin, B., Caroli, B. & Caroli, C. Creep, stick-slip, and dry friction dynamics: experiments and heuristic model. *Phys. Rev. E* **49**, 4973–4988 (1994).
- Dieterich, J. H. & Kilgore, B. Direct observation of frictional contacts: new insights for state-dependent properties. *Pure Appl. Geophys.* **143**, 283–302 (1994).
- Byerlee, J. D. Friction of rock. *Pure Appl. Geophys.* **116**, 615–626 (1978).
- Segall, P. & Rice, J. R. Dilatancy, compaction, and slip instability of a fluid infiltrated fault. *J. Geophys. Res.* **100**, 22155–22171 (1995).
- Gu, J. C., Rice, J. R., Ruina, A. L. & Tse, S. T. Slip motion and stability of a single degree of freedom elastic system with rate and state dependent friction. *J. Mech. Phys. Solids* **32**, 167–196 (1984).
- Beeler, N. M., Tullis, T. E. & Weeks, J. D. The roles of time and displacement in the evolution effect in rock friction. *Geophys. Res. Lett.* **21**, 1987–1990 (1994).
- Blanpied, M. L., Lockner, D. A. & Byerlee, J. D. Fault stability inferred from granite sliding experiments at hydrothermal conditions. *Geophys. Res. Lett.* **18**, 609–612 (1991).

22. Scholz, C. H. The brittle-plastic transition and the depth of seismic faulting. *Geol. Rundsch.* **77**, 319–328 (1988).
23. Shimamoto, T. A transition between frictional slip and ductile flow undergoing large shearing deformation at room temperature. *Science* **231**, 711–714 (1986).
24. Marone, C., Raleigh, C. B. & Scholz, C. Frictional behavior and constitutive modelling of simulated fault gouge. *J. Geophys. Res.* **95**, 7007–7025 (1990).
25. Marone, C. & Scholz, C. The depth of seismic faulting and the upper transition from stable to unstable slip regimes. *Geophys. Res. Lett.* **15**, 621–624 (1988).
26. Sibson, R. H. Fault zone models, heat flow, and the depth distribution of earthquakes in the continental crust of the United States. *Bull. Seismol. Soc. Am.* **72**, 151–163 (1982).
27. Byrne, D. E., Davis, D. M. & Sykes, L. R. Loci and maximum size of thrust earthquakes and the mechanics of the shallow region of subduction zones. *Tectonics* **7**, 833–857 (1988).
28. Hyndman, R. D. & Wang, K. Thermal constraints on the zone of major thrust earthquake failure: the Cascadia subduction zone. *J. Geophys. Res.* **98**, 2039–2060 (1993).
29. Sibson, R. H. Transient discontinuities in ductile shear zones. *J. Struct. Geol.* **2**, 165–171 (1980).
30. Stel, H. The effect of cyclic operation of brittle and ductile deformation on the metamorphic assemblage in cataclases and mylonites. *Pure Appl. Geophys.* **124**, 289–307 (1986).
31. Kanamori, H. & Kikuchi, M. The 1992 Nicaragua earthquake: a slow tsunami earthquake associated with subducted sediments. *Nature* **361**, 714–716 (1993).
32. Irwin, W. P. & Barnes, I. Effects of geological structure and metamorphic fluids on seismic behavior of the San Andreas fault system in central and northern California. *Geology* **3**, 713–716 (1975).
33. Ruff, L. & Kanamori, H. Seismicity and the subduction process. *Phys. Earth Planet. Inter.* **23**, 240–252 (1980).
34. Zoback, M. L. & Zoback, M. D. Crustal stress and intraplate deformation. *Geowissenschaften* **15**, 116–112 (1997).
35. Zoback, M. D. *et al.* New evidence on the state of stress of the San Andreas fault system. *Science* **238**, 1105–1111 (1987).
36. Scholz, C. H. Faults without friction? *Nature* **381**, 556–557 (1996).
37. Scholz, C. H. & Campos, J. On the mechanism of seismic decoupling and back-arc spreading in subduction zones. *J. Geophys. Res.* **100**, 22103–22112 (1995).
38. McCaffrey, R. Statistical significance of the seismic coupling coefficient. *Bull. Seismol. Soc. Am.* **87**, 1069–1073 (1997).
39. Scholz, C. H. & Small, C. The effect of seamount subduction on seismic coupling. *Geology* **25**, 487–490 (1997).
40. Seeber, L., Armbruster, J. G. & Quittmeyer, R. C. *Seismicity and continental subduction in the Himalayan arc* in *Geodynamics Series, V* (eds Gupta & Delany) 215–242 (Am. Geophys. Un., Washington DC, 1981).
41. Amelung, F. & King, G. C. P. Earthquake scaling laws for creeping and non-creeping faults. *Geophys. Res. Lett.* **24**, 507–510 (1997).
42. Nadeau, R., Foxall, W. & McEvilly, T. Clustering and periodic recurrence of microearthquakes on the San Andreas fault at Parkfield, California. *Science* **267**, 503–507 (1995).
43. Bakun, W. H., Stewart, R. M., Bufe, C. G. & Marks, S. M. Implication of seismicity for failure of a section of the San Andreas fault. *Bull. Seismol. Soc. Am.* **70**, 185–201 (1980).
44. Tse, S. & Rice, J. Crustal earthquake instability in relation to the depth variation of friction slip properties. *J. Geophys. Res.* **91**, 9452–9472 (1986).
45. Rice, J. R. Spatio-temporal complexity of slip on a fault. *J. Geophys. Res.* **98**, 9885–9907 (1993).
46. Thatcher, W. Nonlinear strain buildup and the earthquake cycle on the San Andreas fault. *J. Geophys. Res.* **88**, 5893–58902 (1978).
47. Savage, J. C. & Prescott, W. H. Asthenospheric readjustment and the earthquake cycle. *J. Geophys. Res.* **83**, 3369–3376 (1978).
48. Gilbert, L. E., Scholz, C. H. & Beavan, J. Strain localization along the San Andreas fault: consequences for loading mechanisms. *J. Geophys. Res.* **99**, 23975–23984 (1994).
49. Savage, J. C. & Svarc, J. L. Postseismic deformation associated with the 1992 $M_w = 7.3$ Landers earthquake, southern California. *J. Geophys. Res.* **102**, 7565–7577 (1997).
50. Das, S. & Scholz, C. Theory of time-dependent rupture in the earth. *J. Geophys. Res.* **86**, 6039–6051 (1981).
51. Marone, C. J., Scholz, C. H. & Bilham, R. On the mechanics of earthquake afterslip. *J. Geophys. Res.* **96**, 8441–8452 (1991).
52. Heki, K., Miyazaki, S. & Tsuji, H. Silent fault slip following an interplate thrust earthquake at the Japan trench. *Nature* **386**, 595–598 (1997).
53. Dieterich, J. H. A constitutive law for rate of earthquake production and its application to earthquake clustering. *J. Geophys. Res.* **99**, 2601–2618 (1994).
54. Gross, S. & Kisslinger, C. Estimating tectonic stress rate and state with Landers aftershocks. *J. Geophys. Res.* **102**, 7603–7612 (1997).
55. Dieterich, J. H. Nucleation on faults with rate and state-dependent strength. *Tectonophysics* **211**, 115–134 (1992).
56. Ohnaka, M., Kuwahara, Y., Yamamoto, K. & Hirasawa, T. in *Earthquakes Source Mechanics* (eds Das, S., Boatwright, J. & Scholz, C.) 13–24 (AGU Monogr. 37, Am. Geophys. Un., Washington DC, 1986).
57. Scholz, C. H. The critical slip distance for seismic faulting. *Nature* **336**, 761–763 (1988).
58. Marone, C. & Kilgore, B. Scaling of the critical slip distance for seismic faulting with shear strain in fault zones. *Nature* **362**, 618–622 (1993).
59. Abercrombie, R. & Leary, P. Source parameters of small earthquakes recorded at 2.5 km depth, Cajon Pass, southern California: Implications for earthquake scaling. *Geophys. Res. Lett.* **20**, 1511–1514 (1993).
60. Dodge, D. A., Beroza, G. C. & Ellsworth, W. L. Detailed observations of California foreshock sequences: Implications for the earthquake initiation process. *J. Geophys. Res.* **101**, 22371–22392 (1996).
61. Ellsworth, W. L. & Beroza, G. C. Seismic evidence for a seismic nucleation phase. *Science* **268**, 851–855 (1995).
62. Dieterich, J. H. Nucleation and triggering of earthquake slip: effect of periodic stresses. *Tectonophysics* **144**, 127–139 (1987).
63. Heaton, T. H. Tidal triggering of earthquakes. *Bull. Seismol. Soc. Am.* **72**, 2181–2200 (1982).
64. Hill, D. P. *et al.* Seismicity in the western United States triggered by the M7.4 Landers, California, earthquake of June 28, 1992. *Science* **260**, 1617–1623 (1993).
65. Cotton, F. & Coutant, O. Dynamic stress variations due to shear faults in a plane layered medium. *Geophys. J. Int.* **128**, 676–688 (1997).
66. King, G. C. P., Stein, R. S. & Lin, J. Static stress changes and the triggering of earthquakes. *Bull. Seismol. Soc. Am.* **84**, 935–953 (1994).
67. Power, W. L., Tullis, T. E., Brown, S., Boitnott, G. & Scholz, C. H. Roughness of natural fault surfaces. *Geophys. Res. Lett.* **14**, 29–32 (1987).
68. Carlson, J. M. & Langer, J. S. Mechanical model of an earthquake fault. *Phys. Rev. A* **40**, 6470–6484 (1989).
69. Myers, C., Shaw, B. & Langer, J. Slip complexity in a crustal plane model of an earthquake fault. *Phys. Rev. Lett.* **77**, 972–975 (1996).
70. Shaw, B. E. Frictional weakening and slip complexity on earthquake faults. *J. Geophys. Res.* **100**, 18239–18248 (1995).
71. Ben-Zion, Y. & Rice, J. R. Slip patterns and earthquake populations along different classes of faults in elastic solids. *J. Geophys. Res.* **100**, 12959–12983 (1995).
72. Cochard, A. & Madariaga, R. Complexity of seismicity due to highly rate-dependent friction. *J. Geophys. Res.* **101**, 25321–25336 (1996).
73. Ihmlé, P. F. & Jordan, T. H. Teleseismic search for slow precursors to large earthquakes. *Science* **266**, 1547–1550 (1994).
74. Ihmlé, P. F. Monte Carlo slip inversion in the frequency domain: application to the 1992 Nicaragua slow earthquake. *Geophys. Res. Lett.* **23**, 913–916 (1996).
75. Johnson, T. Time-dependent friction of granite: implications for precursory slip on faults. *J. Geophys. Res.* **86**, 6017–6028 (1981).
76. Bak, P., Tang, C. & Wiesenfeld, K. Self-organized criticality. *Phys. Rev. A* **38**, 364–374 (1988).

Acknowledgements. I thank L. Sykes and B. Shaw for comments. This work was partially supported by the US Geological Survey.

Correspondence should be addressed to the author.

A caspase-activated DNase that degrades DNA during apoptosis, and its inhibitor ICAD

Masato Enari*, Hideki Sakahira*, Hideki Yokoyama*, Katsuya Okawa†, Akihiro Iwamatsu† & Shigekazu Nagata*‡

* Department of Genetics, Osaka University Medical School, 2-2 Yamada-oka, Suita, Osaka 565, Japan

† Central Laboratories for Key Technology, Kirin Brewery Co., 1-13-5 Fukuura, Kanazawa, Yokohama, Kanagawa 236, Japan

‡ Osaka Bioscience Institute, 6-2-4 Furuedai, Suita, Osaka 565, Japan

The homeostasis of animals is regulated not only by the growth and differentiation of cells, but also by cell death through a process known as apoptosis. Apoptosis is mediated by members of the caspase family of proteases, and eventually causes the degradation of chromosomal DNA. A caspase-activated deoxyribonuclease (CAD) and its inhibitor (ICAD) have now been identified in the cytoplasmic fraction of mouse lymphoma cells. CAD is a protein of 343 amino acids which carries a nuclear-localization signal; ICAD exists in a long and a short form. Recombinant ICAD specifically inhibits CAD-induced degradation of nuclear DNA and its DNase activity. When CAD is expressed with ICAD in COS cells or in a cell-free system, CAD is produced as a complex with ICAD: treatment with caspase 3 releases the DNase activity which causes DNA fragmentation in nuclei. ICAD therefore seems to function as a chaperone for CAD during its synthesis, remaining complexed with CAD to inhibit its DNase activity; caspases activated by apoptotic stimuli then cleave ICAD, allowing CAD to enter the nucleus and degrade chromosomal DNA.

Surplus cells are removed by apoptosis¹ during mammalian development, and when cytotoxic T cells or natural killer cells kill virally infected or tumour cells, the effector cells induce apoptosis in the target cells². Apoptotic cell death is characterized by accompanying morphological changes, such as cell shrinkage, condensation of nuclei, and loss of microvilli³. The biochemical hallmark of apoptosis is the cleavage of chromosomal DNA into nucleosomal units, which appears to be the final blow in the cell death process⁴⁻⁶. Until now, none of the nuclease(s) responsible for DNA degradation in apoptosis had been identified, although various nucleases such as DNase I, DNase II, and cyclophilins have been proposed as candidates⁷⁻⁹.

Fas ligand, a member of the tumour-necrosis factor family, is expressed in activated T cells and natural killer cells, and induces apoptosis in target cells by binding to Fas, its receptor^{2,10}. As this process occurs in the presence of inhibitors of protein or RNA synthesis¹¹, and even in enucleated cells¹², all the signalling components necessary for apoptosis may already be present in growing cells, and the apoptotic signals may only trigger them. It has been shown that Fas engagement activates a cascade of caspase proteases¹³⁻¹⁸, as found in most apoptotic processes¹⁹. Caspases are synthesized as precursor forms, and an apoptotic signal converts the precursors to mature enzymes, which subsequently cleave other caspases that are downstream in the cascade^{2,20}. Caspases activated by apoptotic signals cleave various cellular substrates such as actin, poly(ADP-ribose) polymerase, fodrin and lamin, which may be responsible for the morphological changes that occur in the cells²¹.

Fas-induced apoptosis also causes DNA degradation¹¹. The cytosolic fraction from Fas-engaged apoptotic cells contain a factor(s) that induces DNA degradation in isolated nuclei, suggesting that a nuclease or its activating molecule(s) is generated in cells during Fas-induced apoptosis^{22,23}. The cytosolic extracts from living cells do not cause DNA degradation in nuclei, but treatment of the extracts with caspase 3 converts them to apoptosis-inducing extracts, suggesting that living cells carry a proform of the molecule(s) that induces DNA degradation¹⁴. We have now purified a protein that can cause DNA degradation in nuclei after treatment with caspase 3.

As this protein has an intrinsic DNase activity, it was designated CAD, for caspase-activated DNase. We also tested for the presence of an inhibitor(s) of DNA degradation and found that the cytosolic fraction from growing, but not from apoptotic, cells contains such an inhibitor (ICAD, for inhibitor of CAD). ICAD was also purified from a mouse T-cell lymphoma. Molecular cloning of mouse CAD complementary DNA revealed that CAD is a protein consisting of 343 amino acids, carrying a nuclear-localization signal at its carboxy terminus. Two different ICAD cDNAs were isolated which coded for a long and a short form; a recombinant ICAD produced in *Escherichia coli* specifically inhibited the DNase activity of CAD, but did not inhibit DNase I or DNase II. Functional CAD protein was produced in COS cells in the presence of ICAD in a cell-free system, CAD was only found in complex with ICAD. These results indicate that CAD exists as an inactive complex with ICAD in living cells. Caspases activated by apoptotic signals cleave ICAD to release CAD, which then enters the nucleus to degrade chromosomal DNA. This mechanism of activation of this apoptotic DNase is remarkably similar to that used to activate the transcription factor NF- κ B.

Identification of CAD and ICAD

We have previously shown that growing cells contain a factor(s) that can be converted by caspase 3 to a form that causes DNA degradation¹⁴. To characterize this factor, the S-100 fraction from mouse T-cell lymphoma was fractionated by DEAE-Sepharose column chromatography. Each fraction was treated with caspase 3, then its ability to cause DNA degradation in nuclei was assayed. As shown in Fig. 1a, the DNA degradation activity eluted as a single peak at about 145 mM NaCl. The activity was not detected without caspase-3 treatment (data not shown). This DNA-degrading activity cleaved naked DNA when plasmid DNA was used as substrate (middle panel, Fig. 1a), and caspase-3-dependent (bottom panel). The S-100 fraction from growing cells was found to inhibit the CAD activity present in lysates from Fas-activated cells. To characterize this CAD-inhibitory activity, we assayed CAD activity in Fas-activated cells in the presence of fractions from the DEAE-Sepharose column. As shown in Fig. 1b, proteins eluting at 170 mM NaCl inhibited CAD-induced DNA fragmentation in nuclei, indicating

that growing cells carry a latent form of CAD and its inhibitor (ICAD) in the cytoplasm.

Purification and properties of ICAD

ICAD was purified by ammonium sulphate fractionation and column chromatography on butyl-Sepharose and phenyl-Superose. When the sample was loaded onto a gel-filtration column, ICAD eluted as a single peak with an apparent relative molecular mass (M_r) of 75K. Physical chemical characterization of partially purified ICAD indicated that it is resistant to heat (90 °C) and denaturant (0.1% SDS). We therefore heated proteins in each fraction from the gel filtration column at 90 °C for 15 min and centrifuged them. The supernatants gave a few bands on SDS-polyacrylamide gel electrophoresis (Fig. 2a) and had ICAD activity (Fig. 2b). To assign ICAD to a specific band, the gel was cut into slices for elution: an eluted 32K protein was found to inhibit CAD-induced DNA degradation in nuclei (Fig. 2c) and its DNase activity (Fig. 2d), indicating that the 32K protein is ICAD; the native form of ICAD is presumably a homodimer, because it behaves as a 75K protein on gel-filtration column chromatography. The overall purification of ICAD from the S-100 fraction was ~17,000-fold, with a yield of 11.8% (Table 1).

Purified ICAD was sequenced and five peptide sequences were obtained. One was found in the mouse EST databases and, using this sequence, two different ICAD cDNAs (ICAD-L and ICAD-S) were isolated from a mouse T-cell lymphoma cDNA library. These cDNAs contained open reading frames of 331 and 265 amino acids, respectively (Fig. 3a). The two cDNAs were identical up to amino-acid position 261, after which the sequences differed, suggesting that the ICAD-L and ICAD-S messenger RNAs are generated through alternative splicing. All five peptide sequences obtained with the purified ICAD were found in the ICAD-S sequence, and its calculated M_r (29.169K) agreed well with that estimated from SDS-polyacrylamide gel electrophoresis. Both ICAD-L and ICAD-S are rich in acidic amino acids, and have isoelectric points of ~4.5. A BLAST homology search in GenBank databases indicated that mouse ICAD-L is highly homologous to human DFF45 (DNA-fragmentation factor 45)²⁴. As shown in Fig. 3b, the two sequences are 76.1% identical, indicating that mouse ICAD-L could be a counterpart of human DFF45.

To test the enzymatic properties of ICAD, ICAD-L and ICAD-S were prepared as glutathione-S-transferase (GST) fusion proteins in *E. coli*. As shown in Fig. 4a, recombinant ICAD inhibited the DNA degradation in nuclei induced by lysate from Fas-activated cells. This inhibition by ICAD was dose-dependent: a few ng of ICAD-L were sufficient to inhibit CAD activity in 20 µg lysate from Fas-activated cells. ICAD-L also efficiently inhibited the DNase activity of CAD (Fig. 4b), but it did not inhibit either DNase I or DNase II (Fig. 4c, d). A similarly specific inhibition against CAD was observed with ICAD-S (data not shown).

Purification and cloning of CAD

Loss-of-function mutations of the Fas gene cause lymphadenopathy and splenomegaly². As cytosolic extracts from the lymph nodes of one such mouse mutant, MRL-*lpr*, have a high level of CAD activity, we used them as starting material for purification of CAD. We have noticed that CAD self-activated during purification and that this activated form is unstable. The S-100 fraction was therefore first treated with an inhibitor of caspase 3 (bio-DEVD-cmk)¹⁴, CAD was then purified by ammonium sulphate fractionation and chromatography on heparin-agarose, resource-Q and resource-S columns. CAD activity was assayed by its ability to induce DNA degradation in intact nuclei or by its DNase activity on plasmid DNA. The crude extracts (S-100) contained constitutively active DNase activity, which was mostly removed by resource-Q column chromatography. After this stage, the caspase-activated DNase activity copurified with the DNA-fragmentation activity in intact nuclei.

As already described, ICAD is a mouse homologue of human DFF45. Caspase 3 cleaves DFF45/ICAD^{24,25}, so we considered that CAD could be bound to ICAD as a complex. We therefore biotinylated the partially purified CAD (resource-S fraction), and treated it with caspase 3, then loaded it onto an affinity column of GST-DFF45. When the bound proteins were eluted from the column with glutathione, western blot analysis of the eluates with horseradish peroxidase-conjugated streptavidin revealed a specific band of M_r 40K in the eluate from the GST-DFF45 affinity column, but not in eluate from the control column (GST-revDFF45) (Fig. 5a). When the sample was loaded onto the DFF45 affinity column without caspase 3 treatment, the 40K protein was not retained on the column. These results indicated that the 40K protein was probably CAD, and that it can bind to human DFF45 or ICAD. The inability of the inactive form of CAD to bind to the affinity column suggested that this form could be a complex of CAD and ICAD, and that treatment with caspase releases CAD from ICAD. As GST-DFF45 carries a thrombin cleavage site²⁶ and DFF45 can be cleaved by caspase 3 (ref. 24), we treated the column with thrombin and caspase 3 and eluted a 40K protein (data not shown, and Fig. 5b). The size of this protein was identical to that eluted by glutathione, suggesting that the protein is not cleaved by caspase 3. No potential cleavage site for caspase 3 was found in the primary sequence of CAD (Fig. 5d).

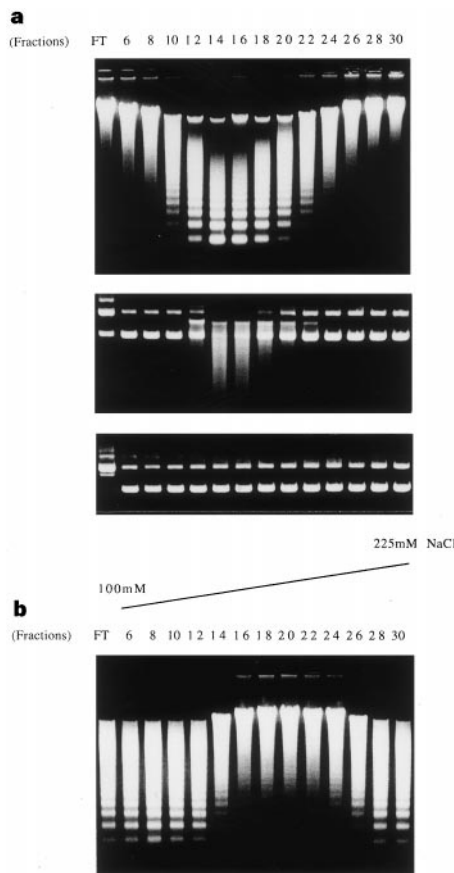


Figure 1 CAD and ICAD activity in extracts from mouse WR19L cells. The S-100 fraction (500 mg protein) from WR19L cells was loaded onto a DEAE-Sepharose column (2.6 × 11 cm), and proteins were eluted with a 50–350 mM NaCl linear gradient. **a**, Aliquots (10 µl) of the indicated fractions were treated with caspase 3, and CAD activity on nuclear (top) or plasmid DNA (middle) determined. DNase activity was also determined without caspase 3 treatment (bottom). **b**, Using 10-µl aliquots, ICAD activity in the indicated fractions was determined using lysates from Fas-activated W4 cells and mouse liver nuclei. FT, flow through.

To prepare enough purified CAD for protein-sequence analysis, 2 mg of the resource-S fraction was activated with caspase 3 and loaded onto the GST–DFF45 affinity column. As the elution with a combination of caspase 3 and thrombin is more efficient than with glutathione, the column was treated with thrombin and caspase 3. Analysis of the eluate by silver staining (Fig. 5b) revealed a band at M_r 40K. Several intense bands smaller than 35K were obtained after thrombin and caspase-3 treatment. The eluate from the affinity column had strong CAD activity. As shown in Fig. 5c, one microlitre of the eluate, which contained less than 1 ng of the 40K protein, completely fragmented nuclear DNA within 2 hours and also digested plasmid DNA. If CAD was loaded onto the affinity column without pretreatment with caspase 3, there was no 40K protein or CAD activity in the eluate (Fig. 5b, c). We concluded that the 40K protein is CAD. A summary of the purification is shown in Table 2: CAD-induced DNA degradation in nuclei was assayed by measuring the amount of histone/DNA complex released from the nuclei. The overall purification of CAD was ~180,000-fold, with a yield of 2.3%.

Five peptide-fragment sequences of the putative CAD protein were obtained, two of which were used to design the degenerate

oligonucleotide primers for polymerase chain reaction (PCR) amplification from a mouse T-cell cDNA library. A fragment of 332 base pairs (bp) was obtained and used as a probe to screen the cDNA library by colony hybridization. One of the positive clones carrying the full-length coding sequence was characterized by nucleotide-sequence analysis. The cDNA (pEF-CAD) contained an open reading frame of 345 amino acids (Fig. 5d). All five peptide sequences obtained from the purified protein could be found in the coding sequence. The amino terminus of the purified CAD was identified as alanine, which appeared at the third position of the coding sequence, suggesting that mouse CAD can undergo post-translational processing at the N terminus. Extra methionine and cysteine residues preceding the N terminus of the mature protein have also been found in actin from various species²⁷. Mature CAD consists of 343 amino acids, with a calculated M_r of 39.333K. Mouse CAD is rich in cysteine residues (12 residues in the mature protein) and is a basic protein with an isoelectric point of 9.7. A BLAST search of the GenBank databases did not reveal any genes homologous to the CAD gene, so CAD is a previously undiscovered gene product. However, the sequence at the C terminus consisting of 15 amino acid residues has the features of a nuclear-localization

Table 1 Purification of murine ICAD

Step	Protein (mg)	Total activity (× 1,000 units)	Specific activity (units μg^{-1})	Purification (fold)	Yield (%)
S-100	1,710.0	20,853.0	12.2		100
DEAE-Sephacrose	125.0	10,975.6	87.8	7	52.6
Ammonium sulphate	62.0	10,812.1	174.3	14	51.8
Butyl-Superose	4.2	3,995.1	951.6	78	19.2
Phenyl-Sephacrose	1.3	3,408.5	2,622.0	215	16.3
Superdex-200	0.12	2,726.3	22,719.5	1,863	13.1
Heat treatment	0.012	2,453.7	204,475.6	16,767	11.8

The S-100 fraction was prepared from 4×10^{11} mouse WR19L cells. ICAD activity was determined with mouse nuclei as described in Methods. One unit of inhibitory activity was defined as the dilution that inhibits half of one unit of CAD activity. The protein concentration of ICAD in the heat-treated sample was estimated from silver staining on SDS-polyacrylamide gel electrophoresis.

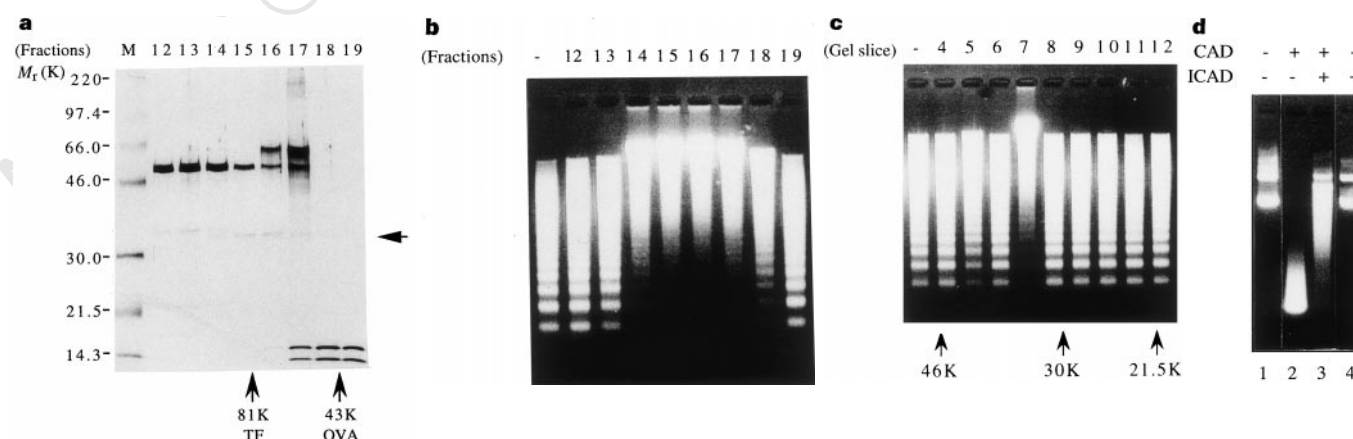


Figure 2 Identification of ICAD. Fractions (1.3 mg protein) from a phenyl-superose column were loaded onto a Superdex-200 column. Fractions were heated and insoluble proteins removed. **a**, Aliquots (5 μl) of the indicated fractions were electrophoresed on a 10–20% gradient polyacrylamide gel and proteins were visualized by silver staining. Migration of marker proteins (Amersham) are shown on the left. The Superdex-200 column was calibrated by using the standard proteins transferrin (TF, 81K) and ovalbumin (OVA, 43K); their elution positions are indicated. ICAD is indicated by the arrow on the right. **b**, Using 5- μl aliquots of the indicated fractions, ICAD activity was determined. CAD activity without ICAD is shown in the first lane (–). **c**, Fractions 14–17 from the Superdex-200 column were pooled, concentrated, and heated. After centrifugation, ~100 ng

protein in the supernatant was separated by electrophoresis on a 10–20% gradient polyacrylamide gel. After slicing up the gel, proteins were eluted and precipitated in cold acetone as described⁴⁴, then dissolved in buffer D containing 6 M guanidine-HCl. The solution was incubated at room temperature for 20 min and dialysed at room temperature against buffer D. Using 5- μl aliquots, ICAD activity was determined. The first lane (–) shows CAD activity without ICAD. **d**, Aliquot (5 μl) of the protein eluate from gel slice 7 in **c** was incubated at 4°C for 30 min with the lysate from Fas-activated W4 cells (20 μg), and the remaining CAD activity was determined with plasmid DNA. Lane 1, plasmid DNA; lane 2, DNA incubated with CAD; lane 3, DNA incubated with CAD and ICAD; lane 4, DNA incubated with ICAD.

signal²⁸, which is in agreement with the ability of CAD to enter the nucleus to degrade DNA.

Expression of CAD *in vitro*

To confirm that our cloned CAD cDNA encodes the protein with CAD activity, a CAD-expression plasmid (pEF-CAD) was introduced into COS cells and cell lysates prepared. As shown in Fig. 6a, untreated lysates did not induce DNA degradation, but when treated with caspase 3, the lysates had CAD activity that was slightly higher (4-fold) than that of empty-vector-transfected cells. As this expression of CAD was low, we considered that either functional CAD could not be synthesized in the absence of ICAD, or that the transfected cells were immediately killed by CAD. To test these possibilities, we introduced the CAD expression plasmid into COS cells together with the expression plasmid for ICAD. Expression of ICAD alone in COS cells inhibited endogenous CAD, and the lysates

had no CAD activity even after treatment with caspase 3. But when CAD was co-expressed with ICAD, the lysates had strong CAD activity. As shown in Fig. 6b, 10–20 ng of total lysate from cells cotransfected with CAD and ICAD efficiently cleaved chromosomal DNA in nuclei and digested plasmid DNA when treated with caspase 3. CAD activity in the transfected cells was ~1,000 times higher than in control cells transfected with empty vector, indicating that our cloned cDNA codes for functional CAD.

To investigate the role of ICAD in the expression of CAD, we synthesized CAD in a cell-free system. Specifically, CAD mRNA was synthesized with T7 polymerase and translated in reticulocyte lysates in the presence or absence of recombinant GST-ICAD. In both cases, CAD was synthesized as a 40K protein (Fig. 7a). CAD made in the absence of ICAD was inactive but was fully functional when synthesized in the presence of ICAD and activated by caspase 3 (Fig. 7b). When ICAD was added to the reaction after CAD



Figure 3 Comparison of amino-acid sequences of mouse ICAD-S and ICAD-L and of mouse ICAD-L and human DFF-45. **a**, The amino-acid residues in ICAD-L that are identical to those of ICAD-S are indicated by a dash. The five peptide sequences obtained from purified ICAD are underlined. Two putative cleavage

sites for caspase 3 are indicated by double underlines. The amino-acid number is shown above the sequence. **b**, Alignment of amino-acid sequences of mouse ICAD-L and human DFF45. The amino-acid sequence of human DFF45 (ref. 24) was aligned with that of mouse ICAD-L. Identical amino acids are in bold.

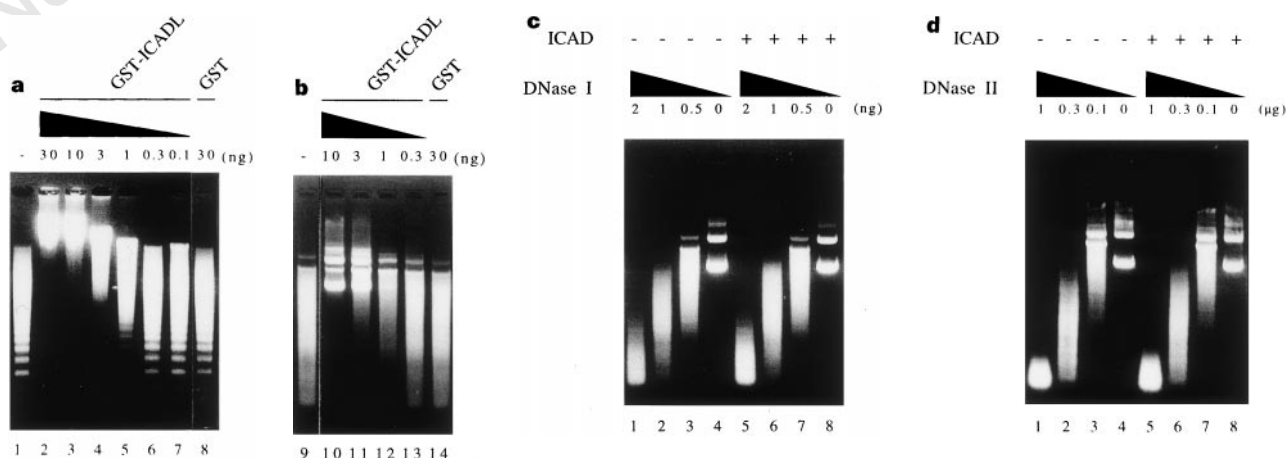


Figure 4 Specific inhibition of CAD activity by recombinant ICAD. **a** and **b**, Cell lysate from Fas-activated W4 cells (20 μg) was preincubated with the indicated amounts of GST-ICAD-L, and the CAD activity was determined with **a**, liver nuclei, **b**, or plasmid DNA. The activity of CAD without preincubation with GST-ICAD-L is shown in lanes 1 and 9. The effect of the control GST protein on CAD activity is

shown in lanes 8 and 14. **c**, **d**, Increasing amounts of DNase I (**c**) or DNase II (**d**) were preincubated at 4°C for 30 min with 10 ng (**c**) or 10 μg (**d**) of GST-ICAD-L. Plasmid DNA (1.0 μg) was added and the incubation continued at 30°C for 2 h, then analysed by agarose gel electrophoresis. Lanes 1–4, DNase alone; 5–8, DNase with ICAD-L.

synthesis, however, no CAD activity was detected (data not shown). These results indicate that functional CAD can only be synthesized in the presence of ICAD, and suggest that newly synthesized CAD is complexed with ICAD. When ICAD tagged with Flag epitope was immunoprecipitated with anti-Flag antibody, the 40K CAD protein coimmunoprecipitated with ICAD (data not shown); treatment of the precipitate with caspase 3 gave full CAD activity (Fig. 7c), confirming that CAD is a DNase that forms a complex with ICAD and is activated by caspase 3.

Discussion

Ever since the discovery that DNA fragmentation occurs during apoptosis⁶, the identity of the DNase responsible has been sought. However, this DNase was labile and scanty, making purification of this enzyme difficult²⁹. Several enzymes have been proposed as candidates, including DNase I, DNase II, cyclophilins, and DNase- γ (refs 7–9, 30), but none of these fulfilled all the criteria for the apoptotic DNase. CAD is a DNase with a high specific activity, comparable to or higher than that of DNase I and DNase II, which can be generated from cell extracts by treatment with caspase 3, and we have also identified its specific inhibitor, ICAD. By using recombinant CAD produced in COS cells, we have shown that

CAD's DNase activity can be inhibited by aurintricarboxylic acid at 30 μ M (M.E. and S.N., unpublished data), a concentration that inhibits DNA degradation in most apoptotic cells^{31,32}. Northern blot analysis indicates that CAD and ICAD mRNAs are ubiquitously expressed in various tissues (data not shown). The amino-acid sequence of CAD contains a nuclear-localization signal. Furthermore, overexpression of ICAD blocks DNA degradation induced by Fas engagement or by treatment with staurosporine²⁵. These results indicate that CAD is the DNase responsible for DNA degradation during apoptosis, although we cannot rule out the possibility that other DNases may be involved in some circumstances.

The inactive form of CAD in non-apoptotic growing cells seems to exist as a complex of CAD with ICAD, because the caspase-activated but not the inactive form of CAD could bind free ICAD, and because CAD synthesized in a cell-free system was recovered as a complex with ICAD. When partially purified CAD (resource-S fractions) was heated to inactivate CAD, the ICAD activity was found in the fractions that contained CAD (data not shown). CAD is a basic protein (pI 9.7), whereas ICAD is an acidic protein (pI 4.5), so these proteins may interact electrostatically. Caspase 3 is activated during apoptosis^{14,19} and cleaves ICAD²⁵, which probably releases CAD from ICAD. When we analysed the inactive and

Table 2 Purification of murine CAD

Step	Protein (mg)	Total activity ($\times 1,000$ units)	Specific activity (units μ g ⁻¹)	Purification (fold)	Yield (%)
S-100	3,150	1,493	0.47		100
Ammonium sulphate	850	960.5	1.13	2	65
Heparin	89.6	1,333.2	14.9	32	90.2
Resource-Q	16.4	944.6	57.6	123	63.9
Resource-S	2.0	560.2	280.0	596	37.9
GST-DFF45 Sepharose	0.0004	34.3	85,800.0	182,553	2.3

The S-100 fraction was prepared from lymph nodes of 14 MRL-*lpr/lpr* mice. To determine CAD activity, 2×10^5 nuclei were incubated with the sample at 30 °C for 2 h in 20 μ l of buffer A containing 150 ng caspase 3. The histone/DNA complex released from nuclei was assayed with a cell death ELISA kit (Boehringer Mannheim). One unit of CAD activity was arbitrarily defined as the activity that gives one optical density unit at 405 nm in the ELISA system. Ten units of this assay roughly correspond to the activity that degrades 1 μ g plasmid DNA at 30 °C for 2 h. The protein concentration of CAD in the eluate from GST-DFF45 Sepharose was estimated from silver staining after SDS-PAGE.

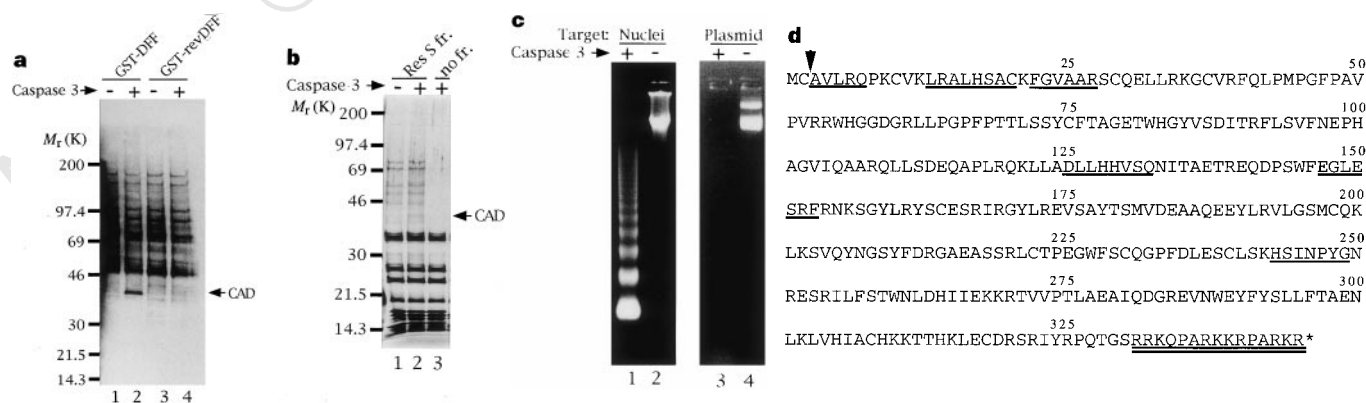


Figure 5 Identification and amino-acid sequence of CAD. **a**, Partially purified CAD (resource-S fraction, 0.1 mg) was biotinylated with biotin-*N*-hydroxysuccinimide ester (Pierce) and treated at 4 °C overnight with (lanes 2 and 4) or without (lanes 1 and 3) 6 μ g ml⁻¹ caspase 3, then the caspase was inactivated with bio-DEVD-cmk. The sample (16 μ g protein) was loaded onto GST-DFF45 affinity beads (10 μ l bed volume) (lanes 1 and 2) or control GST-revDFF beads (lanes 3 and 4). The beads were incubated at 4 °C for 2 h with 20 μ l buffer G containing 5 mM glutathione. Proteins released from the beads were analysed by western blotting with horseradish-peroxidase-conjugated streptavidin and the biotin-streptavidin complex was visualized by an enhanced chemiluminescence system (Renaissance; Dupont NEN). **b**, The resource-S fraction (2 mg protein each) was incubated at 4 °C overnight with (lane 2) or without (lane 1) caspase 3. After inactivating caspase 3 with bio-DEVD-cmk, samples were applied to the GST-DFF45 affinity

column. 5 μ l of fractions (100 μ l) from the affinity column was analysed by electrophoresis on a polyacrylamide gel, and proteins were visualized by silver-staining. Buffer G was subjected to the same procedure; that is, binding to GST-DFF45 affinity column and elution with caspase 3 and thrombin (lane 3). Caspase 3 and thrombin are mostly seen in this lane. **c**, CAD activity in 1- μ l aliquots of fractions (100 μ l) from the affinity column (**b**) was determined with mouse liver nuclei (lanes 1 and 2) or with plasmid DNA (lanes 3 and 4). Lanes 1 and 3, caspase-3-treated sample loaded onto GST-DFF45 affinity column. Lanes 2 and 4, samples loaded on the affinity column without caspase 3 treatment. **d**, Amino-acid sequence of mouse CAD predicted from the cDNA sequence. The N terminus of purified CAD starts with Ala-3. Five peptide sequences obtained from purified CAD are underlined. A putative nuclear-localization signal at the C terminus is indicated by a double underline.

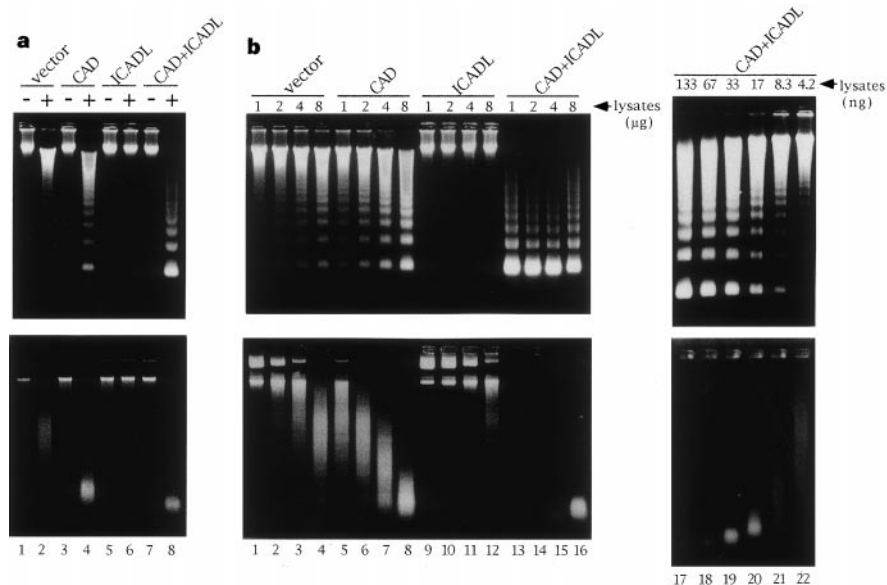


Figure 6 Expression of mouse CAD in COS cells. **a**, COS cells were transfected with pEF-BOS (lanes 1 and 2), pEF-CAD (lanes 3 and 4), or pEF-ICAD-L (lanes 5 and 6), or cotransfected with pEF-CAD and pEF-ICAD-L (lanes 7 and 8), and cell lysates were prepared as described in Methods. Aliquots of the lysates (8 μ g protein) were assayed for CAD activity with nuclei (top) or with plasmid DNA (bottom) in the absence (–) (lanes 1, 3, 5 and 7) or presence (+) (lanes 2, 4, 6 and 8) of caspase 3. **b**, Cell lysates were prepared from COS cells transfected with pEF-BOS (lanes 1–4), pEF-CAD (lanes 5–8), or pEF-ICAD-L (lanes 9–12), or cotransfected with pEF-CAD and pEF-ICAD-L (lanes 13–22). Aliquots from each lysate were assayed for CAD activity with mouse liver nuclei (top) or with plasmid DNA (bottom) in the presence of caspase 3.

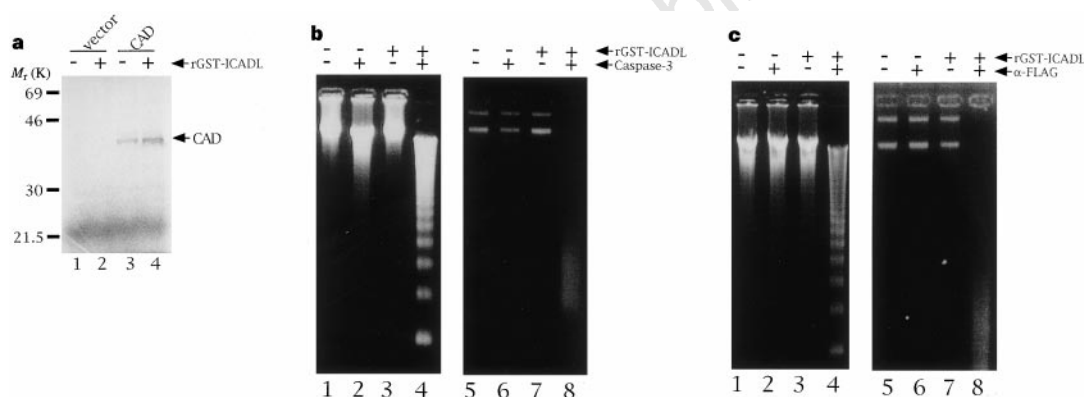


Figure 7 Expression of mouse CAD in a cell-free system, and binding of CAD with ICAD. **a**, The pcDNA-CAD (lanes 3 and 4) or the empty vector (lanes 1 and 2) was subjected to the *in vitro* transcription and translation in the presence of 35 S-Met. The reaction mixtures in lanes 2 and 4 included 160 ng of GST-ICAD-L. After reaction, 5- μ l aliquots were analysed by electrophoresis on a polyacrylamide gel and autoradiographed. **b**, Mouse CAD was synthesized in the *in vitro* transcription and translation system in the absence (lanes 1, 2, 5 and 6) or presence (lanes 3, 4, 7 and 8) of GST-ICAD-L. Using 3 μ l aliquots, CAD activity with

nuclei (lanes 1–4) or plasmid DNA (lanes 5–8) was determined in the absence (lanes 1, 3, 5 and 7) or presence (lanes 2, 4, 6 and 8) of caspase 3. **c**, Mouse CAD was synthesized in the cell-free system in the absence (lanes 1 and 2) or presence (lanes 3 and 4) of GST-ICAD-L. The ICAD-L was immunoprecipitated with anti-Flag antibody (lanes 2, 4, 6 and 8), or control antibody (lanes 1, 3, 5 and 7) and the precipitates suspended in 10 μ l buffer G. Using 3- μ l aliquots, CAD activity with nuclei (lanes 1–4) or plasmid DNA (lanes 5–8) was determined in the presence of caspase 3.

caspase-3-activated forms of CAD by gel filtration, their apparent M_r values were 80K and 40K, respectively (data not shown). This mechanism for the activation of CAD (Fig. 8) is very similar to that used by the nuclear transcription factor NF- κ B (refs 33, 34): NF- κ B is kept as a complex with I κ B in the cytoplasm until various stimuli activate a kinase cascade which leads to the ubiquitin-dependent degradation of I κ B. The degraded I κ B dissociates from NF- κ B, which then enters the nucleus to activate or repress the expression of various genes. In the apoptotic pathway, various stimuli, including cytokines, activate a protease cascade of caspases, which could degrade ICAD and cause its dissociation from CAD; CAD could then enter the nucleus to degrade the chromosomal DNA. NF- κ B carries a nuclear-localization signal, as does CAD, and one of the roles of I κ B is to mask this signal to prevent NF- κ B from entering the nucleus³⁵. ICAD may inhibit not only the DNase activity of CAD, but also its translocation into the nucleus, because most inactive CAD is found in the cytoplasm (data not shown).

Most DNases and proteases are sequestered into lysosomes as soon as they are synthesized. Some proteases, such as caspases in the

cytoplasm, exist in inactive precursor forms which are activated by various stimuli¹⁹. This CAD/ICAD partnership is analogous to that of the bacterial DNases that are encoded by colicin genes and are coordinately expressed with their inhibitors ('immunity proteins'), immediately forming a complex with them³⁶. ICAD works as a specific chaperone, allowing the proper folding of the nascent CAD polypeptide. Chaperones are proteins that prevent incorrect association of, within or between polypeptides during *de novo* protein folding, and also function to guide proteins to cellular organelles while maintaining their unfolded structure³⁷. Although most chaperones are multifunctional and ubiquitous, chaperone-like functions have been identified that are specific for the β -subunit of a potassium channel³⁸ and for one of the subunits of the 20S proteasome³⁹. The subunit of the 20S proteasome is synthesized as a precursor which is processed during the assembly of the proteasomes from their subunits. CAD is synthesized with two extra amino acids at its N terminus, which may be removed by association with ICAD.

Our results described here and in an accompanying Letter²⁵

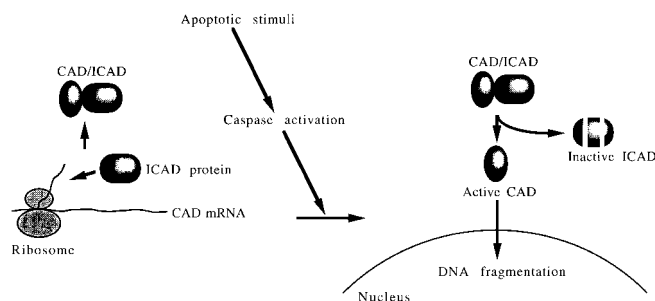


Figure 8 A model for the function of CAD and ICAD in apoptosis. When CAD protein is synthesized, ICAD may bind to the nascent chain of CAD to allow its correct folding. ICAD remains complexed with CAD to inhibit the DNase activity of CAD and to mask its nuclear-localization signal to keep CAD in the cytoplasm. When apoptotic stimuli activate caspases, ICAD is cleaved and, once released, CAD can enter the nucleus where it degrades chromosomal DNA.

establish CAD as the long-sought-after DNase that cleaves chromosomal DNA during apoptosis. The identification and molecular cloning of CAD and ICAD, which link the apoptotic proteases (caspases) to chromosomal DNA degradation, contribute to our understanding of the apoptotic process that originates at the death-factor receptors or from the action of anticancer drugs. □

Methods

Assay for CAD and ICAD. CAD activity was determined by DNA-fragmentation assay using mouse liver nuclei²² or by a DNase activity assay. Nuclei (2×10^5) were incubated at 30°C for 2 h with samples in 20 µl buffer A (10 mM HEPES-KOH, pH 7.0, 50 mM NaCl, 20% (v/v) glycerol, 40 mM β-glycerophosphate, 2 mM MgCl₂, 5 mM DTT) containing 1 mg ml⁻¹ BSA and 150 ng caspase 3. DNA was then extracted from the nuclei and analysed by electrophoresis on a 1.5% agarose gel. In some cases, the released DNA/histone complex was quantified with an ELISA system (kit from Boehringer Mannheim). For DNase assay, nuclei in the mixture were replaced by 1 µg plasmid DNA.

ICAD activity was determined by assaying the inhibition of CAD. The cytosolic fraction from Fas-activated cells was used as a source of activated CAD. Mouse W4 cells were treated at 37°C for 20 min with 0.5 µg ml⁻¹ anti-Fas Jo2 antibody⁴⁰, and the S-100 fraction was prepared as described²². Caspase 3 was inactivated by incubation of the lysate at 4°C for 30 min with 10 µM Ac-DEVD-cho (Peptide Institute). A test sample was added to the active CAD (20 µg of cell extract from Fas-activated W4 cells) in a final volume of 20 µl. After incubation at 4°C for 30 min, the remaining CAD activity was determined using nuclei or plasmid DNA as targets, as described.

Recombinant human DFF45 and mouse ICAD. A cDNA coding for human DFF45 (ref. 24) was isolated from a human KT3 cDNA library¹¹ by PCR. The PCR product (1 kb) was fused to the GST gene of pGEX-2T[128/129]²⁶ in forward and reverse orientations (pGEX-DFF and pGEX-revDFF, respectively). The GST-DFF45 fusion protein was expressed in *E. coli* AD202, and adsorbed to glutathione-sepharose 4B. The beads retained about 1 mg GST fusion protein per ml of wet beads. A GST fusion protein with mouse ICAD was produced by a similar method²⁵.

Purification of ICAD and CAD. All packed columns used for protein purification were from Pharmacia. To purify ICAD, mouse WR19L cells were suspended (2×10^9 cells per ml) in buffer B (10 mM HEPES-KOH, pH 7.0, 50 mM NaCl, 5 mM MgCl₂, 5 mM EGTA, 10% glycerol, and 1 mM DTT) supplemented with a mixture of protease inhibitors (1 mM APMSE, 1 µg ml⁻¹ aprotinin, 1 µg ml⁻¹ leupeptin, and 1 µg ml⁻¹ pepstatin A). Cells were disrupted by three cycles of freezing and thawing in a Dounce homogenizer, accompanied by grinding with a pestle during each thawing cycle. The homogenate was spun at 30,000g for 30 min and at 100,000g for 12 h. The S-100 fraction (1,710 mg) was applied to a DEAE-Sephacel FF column (170 ml bed volume). The column was washed with buffer B, and eluted with a 50–350 mM linear NaCl

gradient. Active fractions were pooled, fractionated with ammonium sulphate at 25 to 50% saturation, and dissolved in buffer C (buffer B without 50 mM NaCl) containing 1 M ammonium sulphate. The sample was applied to a butyl-sepharose-4 FF column (18 ml bed volume) equilibrated with buffer C containing 1 M ammonium sulphate. The column was eluted with an ammonium sulphate linear descending gradient of 1.0 to 0 M. Active fractions were pooled, and ammonium sulphate was added to 0.8 M. After removing insoluble material, the sample was applied to a phenyl-superose HR5/5 column and eluted with buffer C. Active fractions were pooled, and Tween 20 was added to 0.02%. The sample was loaded onto a Superdex-200 HR10/30 gel filtration column equilibrated with buffer D (buffer B containing 0.02% Tween 20). Each active fraction was heated at 90°C for 15 min, and the precipitates were removed by centrifugation at 100,000g for 30 min.

To purify CAD, lymph nodes (~70.4 g) were prepared from MRL-*lpr* mice (4–6 months old; from SLC), and suspended in 244 ml of extraction buffer (12.5 mM PIPES-NaOH, 7.5 mM HEPES-KOH, pH 7.0, 12.5 mM KCl, 37.5 mM NaCl, 2 mM MgCl₂, 5 mM EGTA, 1 mM DTT, 5 mM cytochalasin B, 1 mM PMSF, 1 µg ml⁻¹ leupeptin, 1 µg ml⁻¹ pepstatin A, and 30 mM β-glycerophosphate). Cells were disrupted by three cycles of freezing and thawing, and spun at 100,000g for 2 h. The S-100 fraction was incubated at 4°C for 30 min with 0.1 mM of bio-DEVD-cmk (Peptide Institute) to inactivate caspase 3. The sample was fractionated with ammonium sulphate at 25 to 50% saturation, dialysed against buffer E (10 mM Tris-HCl, pH 8.9, 50 mM NaCl, 20% (v/v) glycerol, 1 mM DTT, 0.1 mM APMSE, 0.15% CHAPS), and applied to a 20 ml HiTrap heparin column. After washing with buffer E containing 150 mM NaCl, proteins were eluted with buffer E containing 500 mM NaCl, and passed through a PD-10 column equilibrated with buffer E. The sample was loaded onto a 6 ml resource-Q column equilibrated with buffer E, and the retained proteins were eluted with a linear 150 to 400 mM NaCl gradient. Active fractions were pooled and passed through a PD-10 column equilibrated with buffer F (20 mM HEPES-KOH, pH 7.0, 20% (v/v) glycerol, 0.1 mM APMSE, 1 mM DTT and 0.15% CHAPS). The sample was loaded onto a 1 ml resource-S column and the retained proteins were eluted with a 15-ml linear NaCl gradient from 0 to 500 mM. Active fractions were pooled and passed through a PD-10 column equilibrated with buffer G (10 mM HEPES-KOH, pH 7.0, 50 mM NaCl, 20% (v/v) glycerol, 40 mM β-glycerophosphate, 2 mM EGTA, 0.1 mM APMSE, 1 mM DTT, 0.02% Tween 20). The sample was incubated at 4°C overnight with 2.6 µg caspase 3, then caspase 3 was inactivated by incubating at 4°C for 30 min with 5 µM bio-DEVD-cmk. GST-DFF45 affinity beads (25 µl bed volume) were added to the mixture, which was rocked at 4°C for 2 h. After washing with buffer G, beads were suspended in 50 µl buffer G. Caspase 3 (2 µg) and thrombin (6 µg) were added to the suspension and incubated at room temperature overnight and at 37°C for 1 h. Beads were removed by centrifugation; proteins released from the beads were collected.

Peptide sequence analysis. The sample was electrophoresed on a 10–20% gradient polyacrylamide gel and blotted onto a polyvinylidene difluoride membrane (Problot, Applied Biosystems). After staining with ponceau S, the immobilized protein (~10 pmol for ICAD, 2–4 pmol for CAD) was reduced, S-carboxymethylated, and digested *in situ* with *Acromobacter* protease I (ICAD)⁴¹ or with *Acromobacter* protease I and Asp-N (for CAD)⁴². Peptides released from the membrane were fractionated by reverse-phase high-performance liquid chromatography with a Wakosil-II AR C18 300A column (Wako Pure Chemical), and sequenced using a protein sequencer (model PPSQ-23, Shimadzu).

Cloning of mouse CAD and ICAD cDNAs. A cDNA library of mouse WR19L cells was constructed (M. Tanaka and S.N., unpublished results). To isolate mouse ICAD cDNA, two oligonucleotides (5'-GATTCGATGGAGGGAC-3' and 5'-GATAAAGCTCAGCTCTGG-3') were designed from EST clone 615712, which encodes one of the peptide sequences of purified ICAD (Fig. 3). A cDNA (476 bp) was amplified by PCR from a WR19L cDNA library, labelled with ³²P and used as a probe to screen the WR19L cDNA library. Screening of 2×10^5 clones by colony hybridization yielded 8 positive clones. Of these, 2 clones carried ICAD-L and 2 clones carried ICAD-S cDNA. Mouse CAD cDNA was cloned from the cDNA library by PCR using the following degenerate primers, whose design was based on the amino-acid sequence of purified mouse CAD: the sense primer was 5'-AA(A/G)TT(C/T)GGIGTIGCIGC-3'; the antisense primer was 5'-TGI(C/G)(A/T)IAC(A/G)TG(A/G)TGIA(A/G)IA(A/G)(A/G)TC-3' (I, inosine). A PCR product (332 bp) was used as a probe to

screen the WR19L cDNA library by colony hybridization. Eight positive clones were obtained out of 1×10^6 clones. Of these, 4 clones carried the full-length coding sequence of mouse CAD. DNA sequencing was done on AlkFred (Pharmacia) or PRIZM 310 (Applied Biosystems) sequencers.

Transfection of COS cells, and *in vitro* transcription and translation.

Plasmid pEF-ICAD-L is a mammalian expression vector of mouse ICAD-L, described in the accompanying Letter²⁵. Monkey COS7 cells were transfected with plasmid DNA by electroporation as described⁴³. After culturing at 37 °C for 48 h, cells were suspended in 0.1 ml buffer G, disrupted by repeated freezing and thawing, and spun to remove cell debris and nuclei. To express CAD in a cell-free system, mouse CAD cDNA was placed under the T7 promoter of the pcDNA1/Amp vector (Invitrogen), and designated pcDNA-CAD. Coupled transcription and translation of CAD was carried out using a TNT *in vitro* transcription/translation kit (Promega) according to the manufacturer's instructions. pcDNA-CAD DNA (1 µg) was incubated at 30 °C for 2 h with 40 µl TNT reagents and 40 µCi ³⁵S-methionine (Amersham) in the presence of 160 ng GST-ICAD-L fusion protein. To immunoprecipitate the ICAD-L/CAD complex, the reaction mixture was diluted to 0.5 ml with buffer G, and 10 µl of anti-flag M2 affinity beads (Kodak; 3 mg IgG per ml of beads) was added. After incubation at 4 °C for 2 h, beads were recovered by centrifugation and suspended in buffer G.

Received 19 November; accepted 1 December 1997.

- Jacobson, M. D., Weil, M. & Raff, M. C. Programmed cell death in animal development. *Cell* **88**, 347–354 (1997).
- Nagata, S. Apoptosis by death factor. *Cell* **88**, 355–365 (1997).
- Wyllie, A. H., Kerr, J. F. R. & Currie, A. R. Cell death: the significance of apoptosis. *Int. Rev. Cytol.* **68**, 251–306 (1980).
- Compton, M. M. A biochemical hallmark of apoptosis: internucleosomal degradation of the genome. *Cancer Metast. Rev.* **11**, 105–119 (1992).
- Wyllie, A. H., Morris, R. G., Smith, A. L. & Dunlop, D. Chromatin cleavage in apoptosis: association with condensed chromatin morphology and dependence on macromolecular synthesis. *J. Pathol.* **142**, 66–77 (1984).
- Wyllie, A. H. Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation. *Nature* **284**, 555–556 (1980).
- Peitsch, M. C. *et al.* Characterization of the endogenous deoxyribonuclease involved in nuclear DNA degradation during apoptosis (programmed cell death). *EMBO J.* **12**, 371–377 (1993).
- Montague, J. W., Hughes, F. J. & Cidlowski, J. A. Native recombinant cyclophilins A, B, and C degrade DNA independently of peptidylpropyl *cis-trans*-isomerase activity. Potential roles of cyclophilins in apoptosis. *J. Biol. Chem.* **272**, 6677–66784 (1997).
- Barry, M. & Eastman, A. Identification of deoxyribonuclease II as an endonuclease involved in apoptosis. *Arch. Biochem. Biophys.* **300**, 440–450 (1993).
- Nagata, S. & Golstein, P. The Fas death factor. *Science* **267**, 1449–1456 (1995).
- Itoh, N. *et al.* The polypeptide encoded by the cDNA for human cell surface antigen Fas can mediate apoptosis. *Cell* **66**, 233–243 (1991).
- Schulze-Osthoff, K., Walczak, H., Dröge, W. & Krammer, P. H. Cell nucleus and DNA fragmentation are not required for apoptosis. *J. Cell. Biol.* **127**, 15–20 (1994).
- Enari, M., Hug, H. & Nagata, S. Involvement of an ICE-like protease in Fas-mediated apoptosis. *Nature* **375**, 78–81 (1995).
- Enari, M., Talanian, R. V., Wong, W. W. & Nagata, S. Sequential activation of ICE-like and CPP32-like proteases during Fas-mediated apoptosis. *Nature* **380**, 723–726 (1996).
- Longthorne, V. & Williams, G. Caspase activity is required for commitment to Fas-mediated apoptosis. *EMBO J.* **16**, 3805–3812 (1997).
- Armstrong, R. C. *et al.* Fas-induced activation of the cell death-related protease CPP32 is inhibited by Bel-2 and by ICE family protease inhibitors. *J. Biol. Chem.* **271**, 16850–16855 (1996).
- Muzio, M. *et al.* FLICE, a novel FADD-homologous ICE/CED-3-like protease, is recruited to the CD95 (Fas/APO-1) death-inducing signaling complex. *Cell* **85**, 817–827 (1996).
- Boldin, M. P., Goncharov, T. M., Goltsev, Y. V. & Wallach, D. Involvement of MACH, a novel MORT1/FADD-interacting protease, in Fas/APO-1- and TNF receptor-induced cell death. *Cell* **85**, 803–815 (1996).
- Henkart, P. A. ICE family protease: mediators of all apoptotic cell death? *Immunity* **4**, 195–201 (1996).
- Fraser, A. & Evan, G. A license to kill. *Cell* **85**, 781–784 (1996).
- Martin, S. & Green, D. Protease activation during apoptosis: death by a thousand cuts. *Cell* **82**, 349–352 (1995).
- Enari, M., Hase, A. & Nagata, S. Apoptosis by a cytosolic extract from Fas-activated cells. *EMBO J.* **14**, 5201–5208 (1995).
- Martin, S. J. *et al.* Cell-free reconstitution of Fas-, UV radiation- and ceramide-induced apoptosis. *EMBO J.* **14**, 5191–5200 (1995).
- Liu, X., Zou, H., Slaughter, C. & Wang, X. DFF, a heterodimeric protein that functions downstream of caspase-3 to trigger DNA fragmentation during apoptosis. *Cell* **89**, 175–184 (1997).
- Sakahira, H., Enari, M. & Nagata, S. Cleavage of CAD inhibitor in CAD activation and DNA degradation during apoptosis. *Nature* **391**, 96–99 (1998).
- Blanan, M. A. & Rutter, W. J. Interaction cloning: identification of a helix-loop-helix zipper protein that interacts with c-Fos. *Science* **256**, 1014–1018 (1992).
- Zakut, R. *et al.* Nucleotide sequence of the rat skeletal muscle actin gene. *Nature* **298**, 857–859 (1982).
- Dingwall, C. & Laskey, R. Nuclear targeting sequences—a consensus? *Trends Biol. Sci.* **16**, 478–481 (1991).
- Cohen, J. J., Duke, R. C., Fadok, V. A. & Sellins, K. S. Apoptosis and programmed cell death in immunity. *Annu. Rev. Immunol.* **10**, 267–293 (1992).
- Shiohara, D., Iwamatsu, A. & Tamura, S. Purification, characterization, and amino acid sequencing of DNase γ from rat spleen. *Arch. Biochem. Biophys.* **346**, 15–20 (1997).
- Batistatou, A. & Green, L. Internucleosomal DNA cleavage and neuronal cell survival/death. *J. Cell Biol.* **122**, 523–532 (1993).
- Mogil, R. *et al.* Role of DNA fragmentation in T cell activation-induced apoptosis *in vitro* and *in vivo*. *J. Immunol.* **152**, 1674–1683 (1994).
- Verma, I., Stevenson, J., Schwarz, E., Van Antwerp, D. & Miyamoto, S. Rel/NF-κB/IκB family: intimate tales of association and dissociation. *Genes Dev.* **9**, 2723–2735 (1995).
- Baldwin, A. The NF-κB and IκB proteins: new discoveries and insights. *Annu. Rev. Immunol.* **14**, 649–681 (1996).
- Beg, A. *et al.* IκB interacts with the nuclear localization sequences of the subunits of NF-κB: a mechanism for cytoplasmic retention. *Genes Dev.* **6**, 1899–1913 (1992).
- Wallis, R. *et al.* *In vivo* and *in vitro* characterization of overproduced colicin E9 immunity protein. *Eur. J. Biochem.* **207**, 687–695 (1992).
- Hartl, F.-U., Hlodan, R. & Langer, T. Molecular chaperones in protein folding: the art of avoiding sticky situations. *Trends Biol. Sci.* **19**, 20–25 (1994).
- Shi, G. *et al.* β-Subunits promote K⁺ channel surface expression through effects early in biosynthesis. *Neuron* **16**, 843–852 (1996).
- Chen, P. & Hochstrasser, M. Autocatalytic subunit processing couples active site formation in the 20S proteasome to completion of assembly. *Cell* **86**, 961–972 (1996).
- Ogasawara, J. *et al.* Lethal effect of the anti-Fas antibody in mice. *Nature* **364**, 806–809 (1993).
- Iwamatsu, A. S-carboxymethylation of proteins transferred onto polyvinylidene difluoride membranes followed by *in situ* protease digestion and amino acid microsequencing. *Electrophoresis* **13**, 142–147 (1992).
- Iwamatsu, A. & Yoshida-Kubomura, N. Systematic peptide fragmentation of polyvinylidene difluoride (PVDF)-immobilized proteins prior to microsequencing. *J. Biochem. (Tokyo)* **120**, 29–34 (1996).
- Suda, T., Takahashi, T., Golstein, P. & Nagata, S. Molecular cloning and expression of the Fas ligand: a novel member of the tumor necrosis factor family. *Cell* **75**, 1169–1178 (1993).
- Hager, D. A. & Burgess, R. R. Elution of proteins from sodium dodecyl sulfate-polyacrylamide gels, removal of sodium dodecyl sulfate, and renaturation of enzymatic activity: results with sigma subunit of *Escherichia coli* RNA polymerase, wheat germ DNA topoisomerase, and other enzymes. *Analyt. Biochem.* **109**, 76–86 (1980).

Acknowledgements. We thank R. V. Talanian for the caspase 3 expression system, M. A. Blanan for pGEX-2T[128/129], and S. Kumagai for secretarial assistance. This work was supported in part by Grants-in-Aid from the Ministry of Education, Science, Sports and Culture in Japan.

Correspondence and requests for materials should be addressed to S.N. (e-mail: nagata@genetic.med.osaka-u.ac.jp). The nucleotide sequence data will appear in the DDBJ, EMBL and Genbank nucleotide sequence databases under the accession number AB009375, AB009376 and AB009377 for mouse ICAD-L, ICAD-S and CAD, respectively.

Discovery of a supernova explosion at half the age of the Universe

S. Perlmutter^{1,2}, G. Aldering¹, M. Della Valle³, S. Deustua^{1,4}, R. S. Ellis⁵, S. Fabbro^{1,6,7}, A. Fruchter⁸, G. Goldhaber^{1,2}, D. E. Groom¹, I. M. Hook^{1,10}, A. G. Kim^{1,11}, M. Y. Kim¹, R. A. Knop¹, C. Lidman¹², R. G. McMahon⁵, P. Nugent¹, R. Pain^{1,6}, N. Panagia¹³, C. R. Pennypacker^{1,4}, P. Ruiz-Lapuente¹⁴, B. Schaefer¹⁵ & N. Walton¹⁶

The ultimate fate of the Universe, infinite expansion or a big crunch, can be determined by using the redshifts and distances of very distant supernovae to monitor changes in the expansion rate. We can now find¹ large numbers of these distant supernovae, and measure their redshifts and apparent brightnesses; moreover, recent studies of nearby type Ia supernovae have shown how to determine their intrinsic luminosities^{2–4}—and therefore with their apparent brightnesses obtain their distances. The >50 distant supernovae discovered so far provide a record of changes in the expansion rate over the past several billion years^{5–7}. However, it is necessary to extend this expansion history still farther away (hence further back in time) in order to begin to distinguish the causes of the expansion-rate changes—such as the slowing caused by the gravitational attraction of the Universe's mass density, and the possibly counteracting effect of the cosmological constant⁸. Here we report the most distant spectroscopically confirmed supernova. Spectra and photometry from the largest telescopes on the ground and in space show that this ancient supernova is strikingly similar to nearby, recent type Ia supernovae. When combined with previous measurements of nearer supernovae^{2,5}, these new measurements suggest that we may live in a low-mass-density universe.

SN1997ap was discovered by the Supernova Cosmology Project collaboration on 5 March 1997 UT, during a two-night search at the Cerro Tololo Interamerican Observatory (CTIO) 4-m telescope that yielded 16 new supernovae. The search technique finds such sets of high-redshift supernovae on the rising part of their light curves and guarantees the date of discovery, thus allowing follow-up photometry and spectroscopy of the transient supernovae to be scheduled¹. The supernova light curves were followed with scheduled R-, I- and some B-band photometry at the CTIO, WIYN, ESO 3.6-m, and INT telescopes, and with spectroscopy at the ESO 3.6-m and Keck II telescopes. (Here WIYN is the Wisconsin, Indiana, Yale, NOAO Telescope, ESO is the European Southern Observatory, and INT is the Isaac Newton Telescope.) In addition, SN1997ap was followed with scheduled photometry on the Hubble Space Telescope (HST).

Figure 1 shows the spectrum of SN1997ap, obtained on 14 March 1997 UT with a 1.5-h integration on the Keck II 10-m telescope. There is negligible ($\leq 5\%$) host-galaxy light contaminating the supernova spectrum, as measured from the ground- and space-based images. When fitted to a time series of well-measured nearby type Ia supernova spectra⁹, the spectrum of SN1997ap is most consistent with a 'normal' type Ia supernova at redshift $z = 0.83$ observed 2 ± 2 supernova-restframe days (~ 4 observer's days) before the supernova's maximum light in the rest-frame B band. It is a poor match to the 'abnormal' type Ia supernovae, such as the brighter SN1991T or the fainter SN1986G. For comparison, the spectra of low-redshift, 'normal' type Ia supernovae are shown in Fig. 1 with wavelengths redshifted as they would appear at $z = 0.83$. These spectra show the time evolution from 7 days before, to 2 days after, maximum light.

Figure 2 shows the photometry data for SN1997ap, with significantly smaller error bars for the HST observations (Fig. 2a) than for the ground-based observations (Fig. 2b and c). The width of the light curve of a type Ia supernova has been shown to be an excellent indicator of its intrinsic luminosity, both at low redshift^{2–4} and at high redshift⁵: the broader and slower the light curve, the brighter

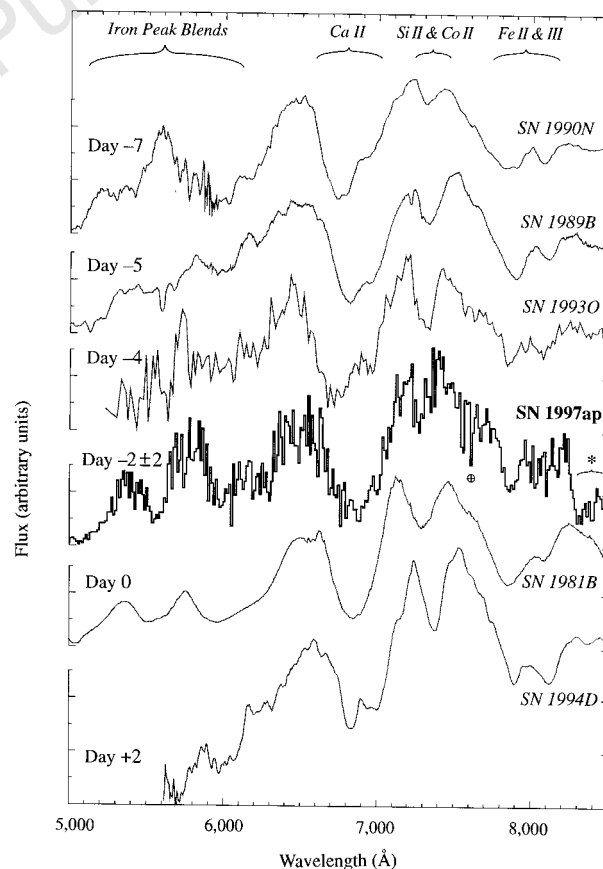


Figure 1 Spectrum of SN1997ap placed within a time sequence of five 'normal' type Ia supernovae. The data for SN1997ap have been binned by 12.5 Å; the time series of spectra of the other supernovae^{17–21} (the spectrum of SN1993O was provided courtesy of the Calán/Tololo Supernova Survey) are given as they would appear redshifted to $z = 0.83$. The spectra show the evolution of spectral features between 7 rest-frame days before, and 2 days after, rest-frame B-band maximum light. SN1997ap matches best at 2 ± 2 days before maximum light. The symbol ⊕ indicates an atmospheric absorption line and * indicates a region affected by night-sky line subtraction residuals. The redshift of $z = 0.83 \pm 0.005$ was determined from the supernova spectrum itself, as there were no host galaxy lines detected.

¹ E. O. Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 50-232, Berkeley, California 94720, USA. ² Center for Particle Astrophysics, University of California, Berkeley, California 94720, USA. ³ Dipartimento di Astronomia, Università di Padova, Vicolo Osservatorio 5, 35122, Padova, Italy. ⁴ Space Sciences Laboratory, University of California, Berkeley, California 94720, USA. ⁵ Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK. ⁶ LPNHE, Universités Paris VI & VII, T33 Rdc, 4, Place Jussieu, 75252 Paris Cedex 05, France. ⁷ Observatoire de Strasbourg, 11, Rue de l'Université, 67000 Strasbourg, France. ⁸ Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. ⁹ Physics Department, Stockholm University, Box 6730, S-11385 Stockholm, Sweden. ¹⁰ European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-85748 Garching bei München, Germany. ¹¹ Physique Corpusculaire et Cosmologie, Collège de France, 11, Place Marcelin-Berthelot, 75231 Paris, France. ¹² European Southern Observatory, Alonso de Cordova, 3107, Vitacura, Casilla 19001, Santiago, 19, Chile. ¹³ Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA; affiliated with the Astrophysics Division, Space Science Department of ESA. ¹⁴ Department of Astronomy, Faculty of Physics, University of Barcelona, Diagonal 647, E-08028 Barcelona, Spain. ¹⁵ Department of Physics, Yale University, 260 Whitney Avenue, JWG 463, New Haven, Connecticut 06520, USA. ¹⁶ Isaac Newton Group, Apartado 321, 38780 Santa Cruz de La Palma, The Canary Islands, Spain.

the supernova is at maximum. We characterize this width by fitting the photometry data to a 'normal' type Ia supernova template light curve that has its time axis stretched or compressed by a linear factor, called the 'stretch factor'^{1,5}; a 'normal' supernova such as SN1989B, SN1993O or SN1981B in Fig. 1 thus has a stretch factor of $s \approx 1$. To fit the photometry data for SN1997ap, we use template U- and B-band light curves that have first been $1+z$ time-dilated and wavelength-shifted ('K-corrected') to the R- and I-bands as they would appear at $z = 0.83$ (see ref. 5 and P.N. *et al.*, manuscript in preparation). The best-fit stretch factor for all the photometry of Fig. 2 indicates that SN1997ap is a 'normal' type Ia supernova: $s = 1.03 \pm 0.05$ when fitted for a date of maximum at 16.3 March 1997 UT (the error-weighted average of the best-fit dates from the

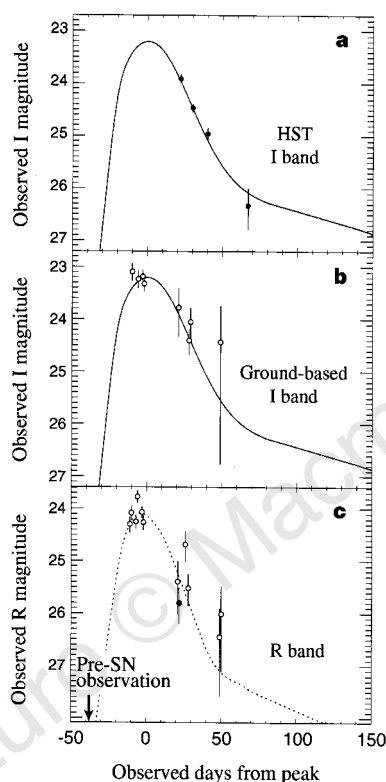


Figure 2 Photometry points for SN1997ap. **a**, As observed by the HST in the F814W filter; **b**, as observed with ground-based telescopes in the Harris I filter; and **c**, as observed with the ground-based telescopes in the Harris R filter (open circles) and the HST in the F675W filter (filled circle); with all magnitudes corrected to the Cousins I or R systems¹³. The solid line shown in both **a** and **b** is the simultaneous best fit to the ground- and space-based data to the K-corrected, $(1+z)$ time-dilated Leibundgut B-band type Ia supernova template light curve²², and the dotted line in **c** is the best fit to a K-corrected, time-dilated U-band type Ia supernova template light curve. The ground-based data was reduced and calibrated following the techniques of ref. 5, but with no host-galaxy light subtraction necessary. The HST data was calibrated and corrected for charge-transfer inefficiency following the prescriptions of refs 23, 24. K-corrections were calculated as in ref. 25, modified for the HST filter system. Correlated zero-point errors are accounted for in the simultaneous fit of the light curve. The errors in the calibration, charge-transfer inefficiency correction and K-corrections for the HST data are much smaller ($\sim 4\%$ total) than the contributions from the photon noise. No corrections were applied to the HST data for a possible $\sim 4\%$ error in the zero points (P. Stetson, personal communication) or for nonlinearities in the WFPC2 response²⁶, which might bring the faintest of the HST points into tighter correspondence with the best-fit light curve in **a** and **c**. Note that the individual fits to the data in **a** and **b** agree within their error bars, providing a first-order cross-check of the HST calibration.

light curve, 15.3 ± 1.6 March 1997 UT, and from the spectrum, 18 ± 3 March 1997 UT).

It is interesting to note that we could alternatively fit the $1+z$ time dilation of the event while holding the stretch factor constant at $s = 1.0^{+0.05}_{-0.14}$ (the best fit value from the spectral features obtained in ref. 10). We find that the event lasted $1+z = 1.86^{+0.31}_{-0.09}$ times longer than a nearby $s = 1$ supernova, providing the strongest confirmation yet of the cosmological nature of redshift^{9,11,12}.

The best-fit peak magnitudes for SN1997ap are $I = 23.20 \pm 0.07$ and $R = 24.10 \pm 0.09$. (All magnitudes quoted or plotted here are transformed to the standard Cousins¹³ R and I bands.) These peak magnitudes are relatively insensitive to the details of the fit: if the date of maximum is left unconstrained or set to the date indicated by the best-match spectrum, or if the ground- and space-based data are fitted alone, the peak magnitudes still agree well within errors.

The ground-based data show no evidence of host-galaxy light, but the higher-resolution HST imaging shows a marginal detection (after co-adding all four dates of observation) of a possible $I = 25.2 \pm 0.3$ host galaxy 1 arcsec from the supernova. This light does not contaminate the supernova photometry from the HST and it contributes negligibly to the ground-based photometry. The projected separation is ~ 6 kpc (for $\Omega_M = 1$, $\Omega_\Lambda = 0$ and $h_0 = 0.65$, the dimensionless cosmological parameters describing the mass density, vacuum energy density and Hubble constant, respectively) and the corresponding B-band rest-frame magnitude is $M_B \approx -17$ and its surface brightness is $\mu_B \approx 21$ mag arcsec⁻², consistent with properties of local spiral galaxies. We note that the analysis will need a final measurement of any host-galaxy light after the supernova has faded, in the unlikely event that there is a very small knot of host-galaxy light directly under the HST image of SN1997ap.

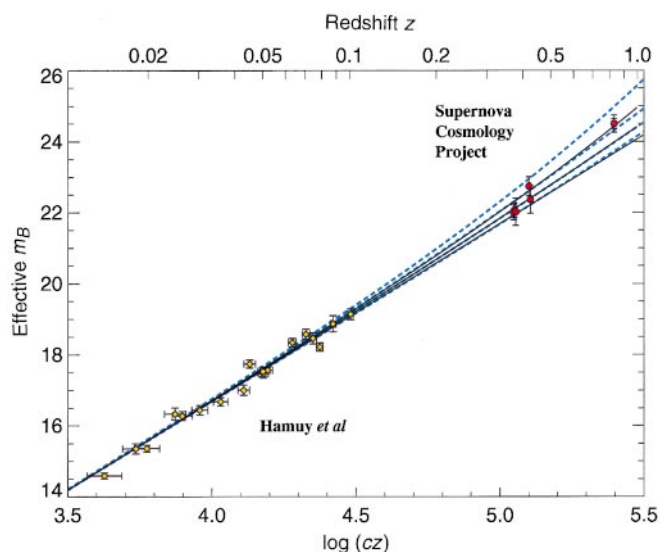


Figure 3 SN1997ap at $z = 0.83$ plotted on the Hubble diagram from ref. 5. Also plotted are the 5 of the first 7 high-redshift supernovae that could be width-luminosity corrected, and the 18 of the lower-redshift supernovae from the Calán/Tololo Supernova Survey that were observed earlier than 5 d after maximum light⁸. Magnitudes have been K-corrected, and also corrected for the width-luminosity relation. The inner error bar on the SN1997ap point corresponds to the photometry error alone, while the outer error bar includes the intrinsic dispersion of type Ia supernovae after stretch correction (see text). The solid curves are theoretical m_B for $(\Omega_M, \Omega_\Lambda) = (0,0)$ on top, $(1,0)$ in middle and $(2,0)$ on bottom. The dotted curves are for the flat-universe case, with $(\Omega_M, \Omega_\Lambda) = (0,1)$ on top, $(0.5,0.5)$, $(1,0)$ and $(1.5,-0.5)$ on bottom.

We compare the K -corrected $R-I$ observed difference of peak magnitudes (measured at the peak of each band, not the same day) to the $U-B$ colour found for 'normal' low-redshift type Ia supernovae. We find that the rest-frame colour of SN1997ap [$(U-B)_{\text{SN1997ap}} = -0.28 \pm 0.11$] is consistent with an unreddened 'normal' type Ia supernova colour, $(U-B)_{\text{normal}} = -0.32 \pm 0.12$ (see ref. 14 and also P.N. *et al.*, manuscript in preparation). In this region of the sky, there is also no evidence for Galactic reddening¹⁵. Given the considerable projected distance from the putative host galaxy, the supernova colour, and the lack of galaxy contamination in the supernova spectra, we proceed with an analysis under the hypothesis that the supernova suffers negligible host-galaxy extinction, but with the following caveat.

Although correcting for $E(U-B) \approx 0.04$ of reddening would shift the magnitude by only one standard deviation, $A_B = 4.8E(U-B) = 0.19 \pm 0.78$, the uncertainty in this correction would then be the most significant source of uncertainty for this one supernova. This is because of the large uncertainty in the $(U-B)_{\text{SN1997ap}}$ measurement, and the sparse low-redshift U -band reference data. HST J -band observations are currently planned for future $z > 0.8$ supernovae, to allow a comparison with the restframe $B-V$ colour, a much better indicator of reddening for type Ia supernovae. Such data will thus provide an important improvement in extinction correction uncertainties for future supernovae and eliminate the need for assumptions regarding host-galaxy extinction. In the following analysis, we also do not correct the lower-redshift supernovae for possible host-galaxy extinction, so any similar distribution of extinction would partly compensate for this possible bias in the cosmological measurements.

The significance of type Ia supernovae at $z = 0.83$ for measurements of the cosmological parameters is illustrated on the Hubble diagram of Fig. 3. To compare with low-redshift magnitudes, we plot SN1997ap at an effective rest-frame B -band magnitude of $B = 24.50 \pm 0.15$, derived, as in ref. 5, by adding a K -correction and increasing the error bar by the uncertainty due to the (small) width-luminosity correction and by the intrinsic dispersion remaining after this correction. By studying type Ia supernovae at twice the redshift of our first previous sample at $z \approx 0.4$, we can look for a correspondingly larger magnitude difference between the

cosmologies considered. At the redshift of SN1997ap, a flat $\Omega_M = 1$ universe is separated from a flat $\Omega_M = 0.1$ universe by almost one magnitude, as opposed to half a magnitude at $z \approx 0.4$. For comparison, the uncertainty in the peak magnitude of SN1997ap is only 0.15 mag, while the intrinsic dispersion amongst stretch-calibrated type Ia supernovae is ~ 0.17 mag (ref. 5). Thus, at such redshifts even individual type Ia supernovae become powerful tools for discriminating amongst various world models, provided observations are obtained, such as those presented here, where the photometric errors are below the intrinsic dispersion of type Ia supernova.

By combining such data spanning a large range of redshift, it is also possible to distinguish between the effects of mass density Ω_M and cosmological constant Λ on the Hubble diagram⁸. The blue contours of Fig. 4 show the allowed confidence region on the Ω_Λ ($\equiv \Lambda/(3H_0^2)$) versus Ω_M plane for the $z \approx 0.4$ supernovae⁵. The yellow contours show the confidence region from SN1997ap by itself, demonstrating the change in slope of the confidence region at higher redshift. The red contours show the result of the combined fit, which yields a closed confidence region in the Ω_M - Ω_Λ plane. This fit corresponds to a value of $\Omega_M = 0.6 \pm 0.2$ if we constrain the result to a flat universe ($\Omega_\Lambda + \Omega_M = 1$), or $\Omega_M = 0.2 \pm 0.4$ if we constrain the result to a $\Lambda = 0$ universe. These results are preliminary evidence for a relatively low-mass-density universe. The addition of SN1997ap to the previous sample of lower-redshift supernovae decreases the best-fit Ω_M by approximately 1 standard deviation compared to the earlier results⁵.

Our data for SN1997ap demonstrate: (1) that type Ia supernovae at $z > 0.8$ exist; (2) that they can be compared spectroscopically with nearby supernovae to determine supernova ages and luminosities and check for indications of supernova evolution; and (3) that calibrated peak magnitudes with precision better than the intrinsic dispersion of type Ia supernovae can be obtained at these high redshifts. The width of the confidence regions in Fig. 4 and the size of the corresponding projected measurement uncertainties show that with additional type Ia supernovae having data of quality comparable to that of SN1997ap, a simultaneous measurement of Ω_Λ and Ω_M is now possible. It is important to note that this measurement is based on only one supernova at the highest ($z > 0.8$) redshifts, and that a larger sample size is required to find a statistical peak and identify any 'outliers'. In particular, SN1997ap was discovered near the search detection threshold and thus may be drawn from the brighter tail of a distribution ('Malmquist bias'). There is similar potential bias in the lower-redshift supernovae of the Calán/Tololo Survey², making it unclear which direction such a bias would change Ω_M .

Several more supernovae at comparably high redshift have already been discovered by the Supernova Cosmology Project collaboration, including SN1996cl, also at $z = 0.83$. SN1996cl can be identified as a very probable type Ia supernova, as a serendipitous HST observation (M. Donahue *et al.*, personal communication) shows its host galaxy to be an elliptical or S0. Its magnitude and colour, although much more poorly constrained by photometry data, agree within uncertainty with those of SN1997ap. The next most distant spectroscopically confirmed type Ia supernovae are at $z = 0.75$ and $z = 0.73$ (ref. 16; these supernovae are awaiting final calibration data). In the redshift range $z = 0.3-0.7$, we have discovered over 30 additional spectroscopically confirmed type Ia supernovae, and followed them with two-filter photometry. (The first sample of supernovae with $z \approx 0.4$ were not all spectroscopically confirmed and observed with two-filter photometry⁵.) These new supernovae will improve both the statistical and systematic uncertainties in our measurement of Ω_M and Ω_Λ in combination. A matching sample of ≥ 6 type Ia supernovae at $z > 0.7$ is to be observed in two filters in Hubble Space Telescope observations due to start on 5 January 1998. SN1997ap demonstrates the efficacy of these complementary higher-redshift measurements in separating

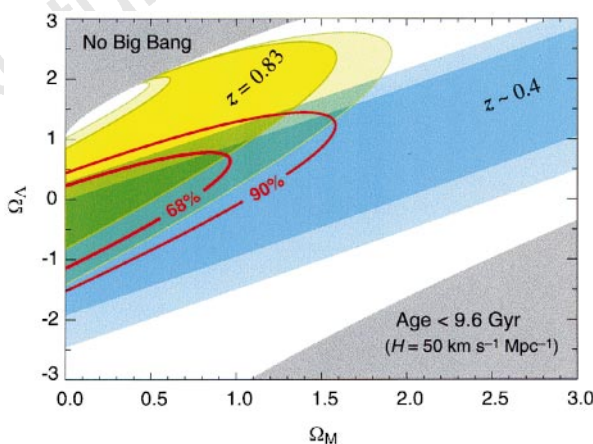


Figure 4 Contour plot of the best fit confidence regions in the Ω_Λ versus Ω_M plane for SN1997ap and the five supernovae at $z \approx 0.4$ (see ref. 5). The 68% (1σ) and 90% confidence regions for (blue shading) the supernovae at $z \approx 0.4$, (yellow shading) SN1997ap at $z = 0.83$ by itself, and (red contours) all of these supernovae taken together. The two labelled corners of the plot are ruled out because they imply: (upper left corner) a 'bouncing' universe with no Big Bang²⁷, or (lower right corner) a universe younger than the oldest heavy elements, $t_0 < 9.6$ Gyr (ref. 28), for any value of $H_0 \geq 50 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

the contribution of Ω_M and Ω_Λ to the total mass-energy density of the Universe.

Received 7 October; accepted 18 November 1997.

1. Perlmutter, S. *et al.* in *Thermonuclear Supernovae* (eds Ruiz-Lapuente, P. *et al.*) 749–763 (Kluwer, Dordrecht, 1997).
2. Phillips, M. M. The absolute magnitudes of Type Ia supernovae. *Astrophys. J.* **413**, L105–L108 (1993).
3. Hamuy, M. *et al.* The absolute luminosities of the Calán/Tololo Type Ia supernovae. *Astron. J.* **112**, 2391–2397 (1996).
4. Riess, A. G., Press, W. H. & Kirshner, R. P. Using Type Ia supernova light curve shapes to measure the Hubble constant. *Astrophys. J.* **438**, L17–L20 (1995).
5. Perlmutter, S. *et al.* Measurements of the cosmological parameters Ω and Λ from the first seven supernovae at $z \geq 0.35$. *Astrophys. J.* **483**, 565–581 (1997).
6. Perlmutter, S. *et al.* *IAU Circ.* No. 6621 (1997).
7. Schmidt, B. *et al.* *IAU Circ.* No. 6646 (1997).
8. Goobar, A. & Perlmutter, S. Feasibility of measuring the cosmological constant Λ and mass density Ω using Type Ia supernovae. *Astrophys. J.* **450**, 14–18 (1995).
9. Riess, A. G. *et al.* Time dilation from spectral feature age measurements of Type Ia supernovae. *Astron. J.* **114**, 722–729 (1997).
10. Nugent, P. *et al.* Evidence for a spectroscopic sequence among Type Ia supernovae. *Astrophys. J.* **455**, L147–L150 (1993).
11. Goldhaber, G. *et al.* in *Thermonuclear Supernovae* (eds Ruiz-Lapuente, P. *et al.*) 777–784 (Kluwer, Dordrecht, 1997).
12. Leibundgut, B. *et al.* Time dilation in the light curve of the distant Type Ia supernova SN 1995K. *Astrophys. J.* **466**, L21–L44 (1996).
13. Bessell, M. S. UBVRI passbands. *Publ. Astron. Soc. Pacif.* **102**, 1181–1199 (1990).
14. Branch, D., Nugent, P. & Fisher, A. in *Thermonuclear Supernovae* (eds Ruiz-Lapuente, P. *et al.*) 715–734 (Kluwer, Dordrecht, 1997).
15. Burstein, D. & Heiles, C. Reddenings derived from H I and galaxy counts—accuracy and maps. *Astron. J.* **87**, 1165–1189 (1982).
16. Perlmutter, S. *et al.* *IAU Circ.* No. 6540 (1997).
17. Leibundgut, B. *et al.* Premaximum observations of the type Ia SN 1990N. *Astrophys. J.* **371**, L23–L26 (1991).
18. Wells, L. A. *et al.* The type Ia supernova 1989B in NGC 3627 (M66). *Astron. J.* **108**, 2233–2250 (1994).
19. Branch, D. *et al.* The type I supernova 1981b in NGC 4536: the first 100 days. *Astrophys. J.* **270**, 123–139 (1983).
20. Patat, F. *et al.* The Type Ia supernova 1994D in NGC 4526: the early phases. *Mon. Not. R. Astron. Soc.* **278**, 111–124 (1996).
21. Cappellaro, E., Turatto, M. & Fernley, J. in *IUE—ULDA Access Guide No. 6: Supernovae* (eds Cappellaro, E., Turatto, M. & Fernley, J.) 1–180 (ESA, Noordwijk, 1995).
22. Leibundgut, B., Tammann, G., Cadonau, R. & Cerrito, D. Supernova studies. VII. An atlas of light curves of supernovae type I. *Astron. Astrophys. Suppl. Ser.* **89**, 537–579 (1991).
23. Holtzman, J. *et al.* The photometric performance and calibration of WFPC2. *Publ. Astron. Soc. Pacif.* **107**, 1065–1093 (1995).
24. Whitmore, B. & Heyer, I. *New Results on Charge Transfer Efficiency and Constraints on Flat-Field Accuracy* (Instrument Sci. Rep. WFPC2 97-08, Space Telescope Science Institute, Baltimore, 1997).
25. Kim, A., Goobar, A. & Perlmutter, S. A generalized K-corrections for Type Ia supernovae: comparing R-band photometry beyond $z = 0.2$ with B, V, and R-band nearby photometry. *Publ. Astron. Soc. Pacif.* **108**, 190–201 (1996).
26. Stiavelli, M. & Mutchler, M. *WFPC2 Electronics Verification* (Instrument Sci. Rep. WFPC2 97-07, Space Telescope Science Institute, Baltimore, 1997).
27. Carroll, S., Press, W. & Turner, E. The cosmological constant. *Annu. Rev. Astron. Astrophys.* **30**, 499–542 (1992).
28. Schramm, D. in *Astrophysical Ages and Dating Methods* (eds Vangioni-Flam, E. *et al.*) 365–384 (Editions Frontières, Gif sur Yvette, 1990).

Acknowledgements. The authors are members of the Supernova Cosmology Project. We thank CTIO, Keck, HST, WIYN, ESO and the ORM—La Palma observatories for a generous allocation of time, and the support of dedicated staff in pursuit of this project; D. Harmer, P. Smith and D. Willmarth for their help as WIYN queue observers; and G. Bernstein and A. Tyson for developing and supporting the Big Throughput Camera which was instrumental in the discovery of this supernova.

Correspondence and requests for materials should be addressed to S.P. (e-mail: saul@lbl.gov).

The inner edge of the accretion disk around a supermassive black hole

B. C. Bromley*, W. A. Miller† & V. I. Pariev‡

* *Theoretical Astrophysics, Harvard-Smithsonian Center for Astrophysics, MS-51, 60 Garden Street, Cambridge, Massachusetts 02138, USA*

† *Theoretical Astrophysics, MS B288, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

‡ *P. N. Lebedev Physical Institute, Leninsky Prospect 53, Moscow 117924, Russia*

Massive black holes are generally thought to exist at the centres of galaxies¹, but an unambiguous identification of a black hole has been impeded by a lack of evidence for the strong-field relativistic effects expected in the vicinity of such an object. Several years ago, a very broad iron emission line was discovered in the active galaxy MCG-6-30-15, indicative of emission from an accretion disk near

the event horizon of a black hole. But this interpretation, based on the line profile, was somewhat model dependent (refs 2–5). Here we present an analysis of the iron-line emission from MCG-6-30-15 that is insensitive to the details (for example, disk thickness and emissivity) of the disk model used. We find that the inner edge of the disk material giving rise to the line is within 2.6 ± 0.3 times the Schwarzschild radius—the event horizon of a non-rotating black hole—at the 95% confidence level. Changes to the disk parameters can only decrease the inner radius of the emitting region, and so we can be confident that we are observing emission from gravitationally bound material in the strong-field region of a supermassive black hole. Moreover, we find that the black hole is rotating at a rate which is $\geq 23 \pm 17\%$ of the theoretical maximum, although this conclusion is model dependent.

The iron K α line from MCG-6-30-15, produced by fluorescence in the presence of a hard X-ray source, has distinct features which are uniquely characteristic of a relativistic accretion disk. The line is peaked near the rest-frame energy of 6.4 keV, with a broad red wing possibly containing a second weaker peak, the analogue of the red horn in the double-peaked, Doppler-shifted profile of a non-relativistic keplerian disk. The red peak in the relativistic case is smoothed by the effect of strong gravitational redshift. The skewness of the profile towards the red, and the sharp blue edge near the rest-frame energy, are evidently unique to relativistic accretion disks³; other mechanisms of line formation, involving collimated jets, outflows, chaotic clouds, and non-disk emitters/absorbers all fail to produce these features. Thus the MCG-6-30-15 iron line is properly interpreted in the context of a relativistic disk.

This iron line, with the kinematical information it contains, offers an opportunity to firmly identify the energy source in the galactic nucleus as a supermassive black hole. A key to this identification is the size of the central object as constrained by the inner radius of disk emission. Fits to the line profile with simple disk models suggest^{2,4,5} that the inner edge of the disk is within a few R_s (where $R_s \equiv 2GM/c^2$ is the Schwarzschild radius) of an object of mass M . Unfortunately, the inner radius inferred from profile fitting is subject to considerable uncertainty, partly from idealizations of the line profile models. These idealizations include a planar (zero-thickness) disk with a time-independent surface emissivity that varies only with radius. Furthermore, profile fitting requires models of the primary hard X-ray continuum, its reflection from the disk, K α emission from species other than iron, contamination from non-disk line emitters, and absorption by intervening gas or dust. Thus it can be difficult to isolate the line emission associated with the disk.

Here we seek to constrain the inner radius of emission without an accurate assessment of the entire line profile. We take advantage of the fact that the extent of the red wing—the minimum frequency at which statistically significant line flux is observed—places a firm limit on the inner radius of the disk. Detectable flux at a particular redshift requires material to be located within some finite radius, and all disk models consistent with this flux must have an inner edge within this radius. Like profile fitting, measurement of the red edge of a line suffers from uncertainties in the continuum emission, which can be very problematical if the red wing has shallow fall-off (as is the case at small radii). Therefore, it is important to choose a conservative measure of the line's extent.

Our analysis is based on the extreme frequency shifts of a line profile as determined by measurable flux in both the red and blue wings. In general, the red edge of the profile comes from emission in the innermost part of the disk where gravitational redshifting is strong, and where material is flowing away from the observer, thereby inducing a Doppler redshift. The gravitational redshift is most important when the inclination angle is low, as the line-of-sight velocity component is relatively small. Within a few Schwarzschild radii, the gravitational redshift is so strong that it dominates over Doppler blueshifts, even from the most rapid line-of-sight

motion toward the observer. Thus, the most significant blueshifts typically arise from emitters well away from the horizon. For example, at a moderate inclination angle of 45° , the maximally blueshifted light originates from emitters at $\sim 5R_s$. Figure 1 illustrates these effects. To make quantitative use of them, we numerically map the relationship between the frequency extrema, the radius and the observed inclination angle of thin disk annuli^{6,7}, and we use this map to obtain bounds on the geometry of disks with finite radial extent.

We examine the frequency extrema of line profiles from MCG-6-30-15 showing the system in three different states, labelled as bright flare (BF), intermediate (Int.) and deep minimum (DM) phases⁴. Significantly, the DM profile is much more skewed and redshifted than the other two. In Fig. 2 we plot the frequency extrema corresponding to the three phases, and superimpose a map of inclination angle and disk radius for a geometrically thin disk. On the basis of extreme redshift alone, the inner edge of the disk in the DM phase is seen to lie within a radius of $2.5 \pm 1.2R_s$. We can reduce this range with data from the Int. and BF phases by using their frequency extrema to constrain the observed inclination angle of the disk. The result is an inclination angle of $29^\circ \pm 4^\circ$, in good agreement with previous estimates^{2,4,5}. (Here we use 95% confident limits throughout.) Our main assumption in this estimate is that the true extreme blueshift of the profile lies within our estimated error bars. This assumption is justified because a sharp blue edge is a feature common to all axisymmetric disk models, and is expected for time-integrated profiles even if the disk has strong asymmetry or inhomogeneities. The angle constraint can then be applied to the DM phase data and we find that the directly detected inner edge of the disk is at $(2.6 \pm 0.3)R_s$.

The frequency extrema analysis of MCG-6-30-15 can be modified to include a more realistic description of the accretion disk than is

given by the thin-disk approximations. The gas in the disk may form a torus with thickness equal to a significant fraction of the radius, and it may be turbulent as it loses angular momentum and flows inwards. Both finite thickness and turbulence will affect the observed frequency extrema⁸. From relativistic calculations using the Novikov–Thorne disk models⁹, we estimate that these effects can broaden the line profile by $\sim 10\%$ at most, assuming that the accretion rate is at the Eddington limit and that the dimensionless viscosity parameter is $\alpha = 0.3$. (For definiteness we assume that the turbulent velocity spectrum is gaussian with root-mean-square velocity proportional to the sound speed in the disk.) The blue edge is affected much more than the red wing, causing both the inclination angle and the inner edge of the disk to be overestimated. In the MCG-6-30-15 data, the turbulent disk model gives an inclination angle of $13^\circ \pm 4^\circ$ and an inner edge of $(2.2 \pm 0.4)R_s$ around a non-rotating black hole. If the hole is rotating at the theoretical maximum rate then the inclination angle is $22^\circ \pm 4^\circ$ while the innermost detectable edge is located at $(2.6 \pm 0.4)R_s$, or even smaller when self-consistent turbulent models¹⁰ are considered.

We emphasize that the radius of the inner disk edge presented here is an upper bound, but one which nonetheless corresponds to directly detected, strongly redshifted flux. The possible effects which could modify this result—finite disk thickness, turbulence, electron scattering of emission-line photons, and a shift in the rest-frame energy from high ionization in the disk (which might occur when accretion rate is near the Eddington limit)—all tend to overestimate the inclination angle and the inner disk edge in our analysis. Therefore the thin-disk approximation gives a conservative bound, and we thus unambiguously identify emitting material within a radius of $(2.6 \pm 0.3)R_s$ around the supermassive black hole in MCG-6-30-15.

The observed X-ray emission from MCG-6-30-15 may reveal

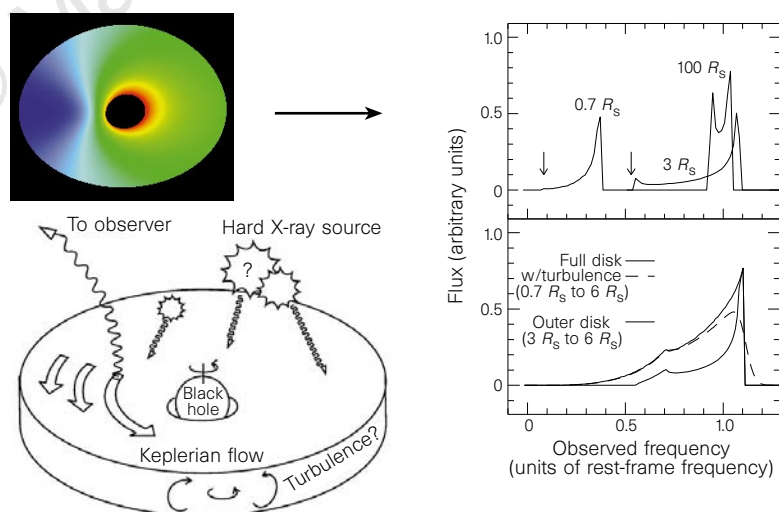


Figure 1 Line emission from an accretion disk around a black hole. The diagram at bottom left indicates the dominant influences on line emission: the source of hard X-rays which cause fluorescence; the velocity flows in the disk which generate Doppler shifts; gravitational redshifting and refocusing of the emitted light. The image at top left shows photon frequency (represented by colour) across the face of the disk as seen by a distant observer⁷. Here the geometrically thin disk lies in the equatorial plane of a maximally rotating black hole and has a radial extent of $(0.7\text{--}6)R_s$; the system is observed at a 45° inclination angle. Emission from narrow rings of disk material at large radii show line profiles—flux versus observed frequency ν —with distinct Doppler peaks from the orbiting emitters, whereas at smaller radii the profiles are more skewed and shifted as a result of strong gravity (upper right plot; arrows indicate extreme redshifts). The

line profile of the full disk (lower-right plot, solid thick line) and that of the outer part between $3R_s$ and $6R_s$ (solid line) are shown for the case where the emissivity of each thin ring varies as a power law in radius with index -2.5 . The red wing of these profiles comes from the inner edge of the disk to the right of the black hole in the image at top left; the blue peak comes from material near the leftmost outer edge. The profile of a turbulent disk (dashed line) is smoothed by random motions within the disk; a Novikov–Thorne disk model with a large amount of turbulence ($\alpha = 0.3$; accretion at the Eddington limit) maximizes the effect. The red edge of the profile is unchanged, but the blue edge is shifted to higher frequencies, causing modest overestimates of both inclination angle and inner radius in thin-disk models.

properties of the central black hole such as its mass and angular momentum. Estimates of the mass are available from the energy released in the galactic nucleus ($M \geq 10^7$ solar masses for MCG-6-30-15), whereas rotation may be constrained by observing the line emission from sources at radii less than $3 R_s$, the marginally stable orbit around a non-rotating black hole. Indeed, previous assessments of the line profile suggest emission from within $3 R_s$ and hence black-hole rotation⁴; a quantitative estimate⁵, made on basis of profile-fitting to a thin-disk model with a fixed emissivity function, suggests a spin $>94\%$ of the theoretical maximum. However, bounds on rotation are possible only if the unstable disk component is of limited emissivity; in some circumstances the disk can be emissive below $3 R_s$ even if the black hole has no rotation¹¹. The unstable part of the disk is then in free fall and it might contribute to the line flux if the accretion rate is high enough to keep the disk optically thick, and if the illuminating X-ray flux is not so great as to completely ionize the iron. On the other hand, if either of these conditions were not met, or if the X-ray production were local (perhaps from magnetic flares caused by differential rotation of the disk), then the observed inner edge would constrain the black-hole rotation. In the case of MCG-6-30-15, this latter interpretation of our results gives a spin of $23 \pm 17\%$ of the theoretical maximum.

We have applied the frequency extrema analysis to the X-ray iron line observed in other Seyfert galaxies. In one sample of 14 Seyfert 1 galaxies¹², half showed very broad profiles; the inferred inclination

angles are $<45^\circ$ and bounds on the inner radii are between $5 R_s$ and $10 R_s$, similar to MCG-6-30-15 in its intermediate phase. A sample of Seyfert 2 galaxies¹³ also show emission lines, but the individual spectra are too noisy to fit line profiles. However, a composite spectrum for the whole sample, which has a higher signal-to-noise ratio, does show a significant broad emission feature. We find a characteristic inner radius bound near $10 R_s$, and an inclination angle which is $<45^\circ$, providing some quantitative support for the assertion¹³ that galaxies in this sample are not more highly inclined than Seyfert 1 galaxies (in contrast with Seyfert unification models based on orientation). Although the inner radius bounds for galaxies in both the Seyfert 1 and Seyfert 2 samples are not as tight as for MCG-6-30-15, they are further evidence of material in the strong-field regime of General Relativity. As X-ray data are accumulated from more objects and with greater precision (by satellite experiments such as AXAF and XMM), we may begin to focus on properties of supermassive black holes. The issue of black-hole rotation in particular will certainly gain much attention, not only for its impact on theories of the formation of active galactic nuclei but also because it offers the possibility of detecting the frame-dragging effect, the gravitational analogue of magnetism in General Relativity. $\square$

Received 15 September; accepted 25 November 1997.

1. Kormendy, J. & Richstone, D. Inward bound—the search for supermassive black holes in galactic nuclei. *Annu. Rev. Astron. Astrophys.* **33**, 581–624 (1995).
2. Tanaka, Y. *et al.* Gravitationally redshifted emission implying an accretion disk and massive black hole in the active galaxy MCG-6-30-15. *Nature* **375**, 659–661 (1995).
3. Fabian, A. C. *et al.* On broad iron K-alpha lines in Seyfert 1 galaxies. *Mon. Not. R. Astron. Soc.* **277**, L11–L15 (1995).
4. Iwasawa, K. *et al.* The variable iron K emission in MCG-6-30-15. *Mon. Not. R. Astron. Soc.* **282**, 1038–1048 (1996).
5. Dabrowski, Y., Fabian, A. C., Iwasawa, K., Lasenby, A. N. & Reynolds, C. S. The profile and equivalent width of the X-ray iron emission-line from a disk around a Kerr black hole. *Mon. Not. R. Astron. Soc.* **288**, L11–L15 (1997).
6. Cunningham, C. T. The effects of redshifts and focusing on the spectrum of an accretion disk around a Kerr black hole. *Astrophys. J.* **202**, 788–802 (1975).
7. Bromley, B. C., Chen, K. & Miller, W. A. Line emission from an accretion disk around a rotating black hole: toward a measurement of frame dragging. *Astrophys. J.* **475**, 57–64 (1997).
8. Pariev, V. I. & Bromley, B. C. Line emission from an accretion disk around a black hole: effects of disk structure. *Astrophys. J.* (submitted).
9. Novikov, I. D. & Thorne, K. S. in *Black Holes* (eds DeWitt, C. & DeWitt, B. S.) 343–450 (Gordon & Breach, New York, 1973).
10. Riffert, H. & Herold, H. Relativistic accretion disk structure revisited. *Astrophys. J.* **450**, 508–511 (1995).
11. Reynolds, C. S. & Begelman, M. C. Iron fluorescence from within the innermost stable orbit of black hole accretion disks. *Astrophys. J.* **488**, 109–118 (1997).
12. Nandra, K., George, I. M., Mushotzky, R. F., Turner, T. J. & Yaqoob, T. ASCA observations of Seyfert 1 galaxies: II. Relativistic iron K α emission. *Astrophys. J.* **477**, 602–622 (1997).
13. Turner, T. J., George, I. M., Nandra, K. & Mushotzky, R. F. ASCA observations of type-2 Seyfert galaxies. III. Orientation and X-ray absorption. *Astrophys. J.* (in the press).

Acknowledgements. We thank S. Jaffe for artwork assistance. The Cray supercomputer used in this work was provided by NASA Offices of Space Sciences, Aeronautics, and Mission to Planet Earth. B.C.B. was supported by the National Science Foundation; W.A.M. and V.I.P. were supported by the US Department of Energy through the LDRD programme at Los Alamos National Laboratory.

Correspondence and requests for materials should be addressed to B.C.B. (e-mail: bbromley@cfa.harvard.edu).

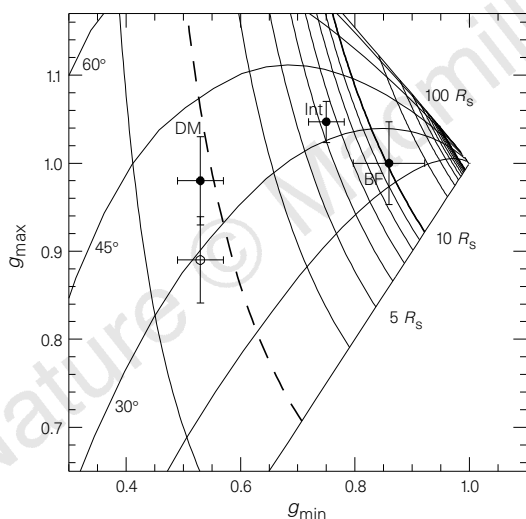


Figure 2 The plane of minimum and maximum redshifts of an emission line, expressed as ratios $g_{\min}$ and $g_{\max}$ of observed frequency to the rest-frame frequency. Superimposed are contours of constant radius and inclination angle which correspond to frequency extrema and geometry of thin rings of material in a geometrically thin disk. The data points correspond to ASCA observations of MCG-6-30-15⁴, and show the frequency extrema as calculated by testing the hypothesis that binned fluxes are purely from continuum emission. An extremum is identified at the frequency where the acceptance level of this hypothesis crosses 0.1% (that is, 99.9% confidence that the flux is above the expected continuum level). The error bars reflect the bin widths. The filled circles are from the deep minimum (DM), intermediate (Int.) and bright flare (BF) data sets. The empty circle is the minimum redshift from the DM data translated so that it is consistent with the inclination angle from the Int. and BF data. We note that this map was produced by assuming that all the material in the thin rings are in circular orbits; below $3 R_s$, the innermost stable orbit of a non-rotating black hole, the spin of the black hole is adjusted so that the frequency extrema at each radius corresponds to the minimum stable orbit. This way we can associate the contours of constant radius with the black-hole spin required to stabilize that radius.

Evidence for a Bose–Einstein condensate in liquid ^4He from quantum evaporation

Adrian F. G. Wyatt

Department of Physics, University of Exeter, Exeter EX4 4QL, UK

Bose–Einstein condensation (BEC) is a purely quantum phenomenon whereby a macroscopic number of identical atoms occupy the same single-particle state¹. Interest in this phenomenon has grown considerably following the direct demonstration of BEC in low-density gases of alkali metal atoms^{2–4}. It is therefore worth reconsidering the case of liquid ^4He , which is generally accepted to

have such a condensate⁵, but for which similarly direct evidence is lacking⁶. Nevertheless, theoretical models that depend on the existence of a condensate have proved successful at explaining many of the properties of this system^{7–9}, and BEC is considered to underlie the striking phenomena of superfluidity and quantized vorticity observed in liquid ⁴He. So the current issue is not whether there is a condensate in this system, but how to demonstrate its existence in a clear and simple way. Here I argue that an earlier measurement¹⁰ of evaporation from liquid ⁴He caused by a collimated beam of phonons provides such a demonstration. The calculated angular distribution of evaporated atoms agrees well with that measured if it is assumed that the atoms initially had zero momentum parallel to the surface of the liquid—this is to be expected if the atoms originate from a condensate. This process of quantum evaporation also opens the possibility for creating beams of phase-coherent atoms of short wavelength.

Quantum evaporation by phonons is analogous to the photoelectric effect^{11,12}. A phonon incident on the free surface of liquid ⁴He is annihilated and an atom is released into the vacuum above the liquid with conservation of momentum parallel to the surface. So measuring the momentum of the phonon and the evaporated atom enables the momentum of the atom in the liquid to be inferred. In an ideal gas of bosons all the particles will be in a zero-momentum state, that is, wavevector $\mathbf{k} = 0$ at $T = 0$. As the temperature is raised the fraction of atoms in the $\mathbf{k} = 0$ state decreases and drops to essentially zero at the BEC temperature. In an interacting and homogeneous system of bosons, the particles have less than unit probability of being in the $\mathbf{k} = 0$ state even at zero temperature¹³. However, a finite condensate fraction, with $\mathbf{k} = 0$ in the bulk or $\mathbf{k}_\parallel = 0$ at a surface, are the signatures of BEC.

The quantum evaporation results show that parallel momentum is conserved between just the phonon and the evaporated atom; this implies that the atom was in a state of zero parallel momentum before it was evaporated¹⁴. As quantum evaporation probes the surface region of the liquid, this indicates at least that there is a Bose–Einstein condensate at the surface of liquid ⁴He as predicted^{15,16}. As yet, no evaporation from the non-condensate fraction has been detected but it is expected to be small¹⁷.

The momentum distribution of the atoms in liquid ⁴He at $T = 0$ is expected to show a considerable range of momenta, as well as a delta function at $\mathbf{k} = 0$ due to the condensate, because of the high number-density of the atoms. The momentum arises from the zero-point motion which is due to the wavefunction of a ⁴He atom having to be zero at neighbouring and other atoms. This is indeed the case¹⁸, and can be seen in the schematic graph of momentum distribution, $n(k)$ (Fig. 1 inset). We see that $0 < k < 2 \text{ \AA}^{-1}$. The average kinetic energy/ k_B per atom is $\sim 14 \text{ K}$, at $T = 0 \text{ K}$ (refs 19, 20) and the corresponding root-mean-square wavevector is $1.2 \text{ \AA}^{-1}$. In a classical liquid the atoms have kinetic energy due to thermal motion, with an average energy $\sim 3/2 k_B T$ per atom. So at the temperature at which most liquids exist, the thermal kinetic energy of an atom is much larger than the zero point kinetic energy; the momentum distribution corresponds to this thermal energy.

Neutron scattering fails to resolve the delta function in $n(k)$ at $\mathbf{k} = 0$, even at temperatures $T \ll T_\lambda$ (the onset temperature for superfluidity) where the condensate is expected to be largest, because the Final States Effect (FSE) and the longitudinal momentum resolution of the spectrometers significantly broaden the spectrum⁶. Nevertheless the neutron-scattering results have been compared with theoretical calculations^{20–22} of the single-atom momentum distribution in liquid ⁴He (ref. 6). The models show a finite condensate containing 9% of the atoms at $T = 0$. To enable comparison with experiment, the model results are broadened by the instrumental resolution and the FSE, which are separately known. The comparison is excellent, showing that the neutron data is consistent with a condensate fraction of $\sim 9\%$ at $T = 0$ and

which decreases to zero at T_λ . However, it has been emphasized that this does not prove the existence of a condensate⁶.

There is a significant difference between evaporating an atom from a classical liquid and from a quantum liquid. An atom will leave a classical liquid if its thermal energy is greater than the chemical potential, μ , of the atom in the liquid (μ being relative to the isolated free-atom state). As $-\mu \gg k_B T$, only atoms in the high-energy tail of the spectrum have enough energy to escape the liquid. An atom that does escape has a kinetic energy, E_v^{KE} , which reflects its thermal kinetic energy in the liquid, E_l^{KE} ; that is, $E_v^{\text{KE}} = E_l^{\text{KE}} + \mu$.

In a quantum liquid at $T = 0$ there is no evaporation. But if a phonon or roton of high enough energy is injected into the liquid and reaches the free surface, it can eject an atom. The excitation is annihilated and the energy, $\hbar\omega$, goes in overcoming the binding energy and giving the free atom kinetic energy $p_v^2/2m_4$ (refs 10–12), that is

$$\hbar\omega = -\mu + p_v^2/2m_4 \quad (1)$$

Here $\mu = -L(0)$, the latent heat per atom at $T = 0$, and m_4 is the mass of a ⁴He atom. The zero-point energy of the atoms in the liquid has a role in determining the latent heat, but the zero-point energy of an individual atom does not contribute to the kinetic energy of that atom when it is evaporated. This can be seen by considering the energy of the liquid ⁴He in its ground state but which contains a phonon of energy $\hbar\omega$. The total energy is $n\mu + \hbar\omega$ where n is the number of ⁴He atoms in the liquid. When the phonon is annihilated and an atom evaporated, the energy is then $(n-1)\mu + p_v^2/2m_4$. For large n the chemical potential is not changed by the evaporation process as the liquid is in its ground state. (This is different from an isolated liquid at a finite temperature where evaporation decreases T and increases μ .) As energy is conserved, we see that $p_v^2/2m_4 = \mu + \hbar\omega$ which is independent of the dynamics of any individual atom. So, for example, if mono-energetic phonons of sufficient energy were injected into the liquid, then mono-energetic atoms would be created in free space. The energy relationship (equation

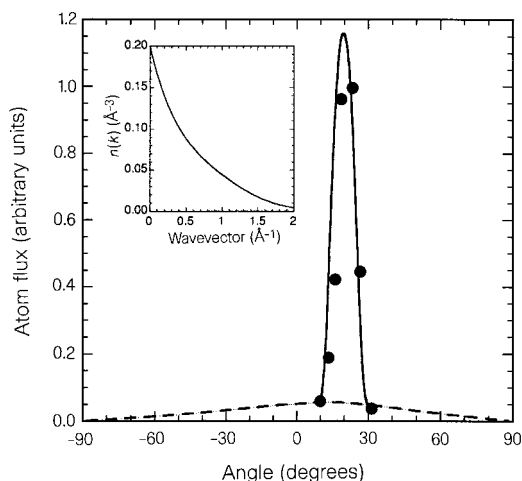


Figure 1 Properties of ⁴He. Main figure, the measured angular distribution of atoms quantum-evaporated from liquid ⁴He by phonons of energy $\sim 10.15 \text{ K}$ at an angle of incidence 25° to the normal; data from ref. 10 (filled circles). The solid line is the result of modelling with the assumption that the atoms had zero momentum in the liquid. The systematic angular difference between the measurements and the model is not significant. The dashed line is the model result for the non-condensate fraction with the momentum distribution shown in the inset. Inset, a schematic graph of the momentum distribution, $n(k)$, for ⁴He atoms in liquid ⁴He at $T \rightarrow 0$, based on ref. 21. It is slightly narrower than the measurements in ref. 18. Only the non-condensate fraction is shown; the condensate fraction would be a delta function at wavevector $k = 0$ representing $\sim 9\%$ of the atoms.

(1)) for quantum evaporation has been verified for phonons and rotons in liquid ^4He by injecting them from a pulsed heater and measuring their time of flight by detecting the atoms with a superconducting bolometer. It is independent of the angle of evaporation over a wide range of angles¹⁰. (The model results shown in Fig. 6b of ref. 10 are superseded by those shown here in Fig. 1, as we now understand that the high-frequency phonon creation process produces a narrow angular beam of phonons^{23,24}.)

Although the range of the zero-point energy of the atoms in liquid ^4He does not show up in quantum evaporation, this is not so for the zero-point momentum which potentially has a strong influence on the atoms evaporated. We can take the free surface of the liquid to be smooth on an atomic scale. This translational invariance leads to the conservation of the component of momentum which is parallel to the surface, that is, the sum of the components of parallel momentum of the atom and phonon in the liquid equals that of the evaporated atom. If the injected phonon interacts with an atom at the interface which has momentum $\mathbf{p}_i$ and so creates a free atom with momentum $\mathbf{p}_v$ then

$$\hbar\mathbf{q}_{\parallel} + \mathbf{p}_{i,\parallel} = \mathbf{p}_{v,\parallel} \quad (2)$$

where $\hbar\mathbf{q}$ is the momentum of the phonon, and the subscript ' $\parallel$ ' denotes the component parallel with the surface. From equation (1) we see that $|\mathbf{p}_v|$ is given by $p_v^2/2m_4 = \hbar\omega - L(0)$. In the direction perpendicular to the surface plane, momentum is not conserved within the phonon-atom system and the liquid recoils as a whole. This conserves momentum overall and the recoiling liquid does not take away any energy because of its high mass.

As the modulus of momentum $|\mathbf{p}_v|$ is determined by the phonon, the effect of equation (2) is just to change the direction of the evaporated atom. All atoms initially with $\mathbf{p}_{i,\parallel} = 0$ evaporate in a unique direction for a given phonon, but atoms with non-zero values of $\mathbf{p}_{i,\parallel}$ will evaporate in different directions. For a beam of mono-energetic phonons evaporating atoms from the liquid with a spectrum of initial momenta $\mathbf{p}_{i,\parallel}$, we would expect evaporated atoms with a wide angular distribution but all having the same kinetic energy. So we see that such a quantum evaporation measurement distinguishes between evaporation from the condensate and non-condensate fractions. For the former we expect a unique angle of evaporation, whereas for the latter we expect a broad angular distribution of evaporated atoms.

An experiment to look for these two different contributions to the evaporated atoms is best done with phonons, as these excitations can be prepared in a beam with a narrow angular width ($\pm 3.5^\circ$) and a narrow energy range, centred at 10.15 K. This felicitous phonon beam arises from the metastable state for phonons with $\omega \geq \omega_c$, the four phonon decay cut-off energy ($\omega_c = 10.0$ K and the corresponding wavevector $q_c = 0.55 \text{ \AA}^{-1}$ at zero pressure) and the way that high-frequency phonons are created by up-scattering from injected low-energy phonons^{23,24}.

The measured angular distribution of atoms for the phonon beam at an angle of incidence of 25° to the normal, at an ambient temperature of 100 mK (ref. 10) is shown in Fig. 1. It is a narrow angular distribution essentially at the angle expected for $\mathbf{p}_{i,\parallel} = 0$. The solid line in Fig. 1 is a model calculation using our best knowledge of the phonon beam and the experimental geometry, and assumes that the atoms have zero momentum in the liquid ^4He . There is a systematic difference of 2° between the data and the model, but this is not significant as the absolute values of angles to the horizontal are only known to within a few degrees. Otherwise the data and the model agree well, showing that the angular width of the atom beam is entirely due to the angular width of the phonon beam and the geometry of the experiment. Incidentally, this also shows that these atoms have not undergone any inelastic processes, such as ripplon production, as these processes would increase the angular width.

The above result suggests that these evaporated atoms are from a condensate with $\mathbf{p}_{i,\parallel} = 0$; we cannot say whether or not there is evaporation from the non-condensate fraction. Evaporated atoms from the non-condensate would be distributed over the entire half-space above the free surface which means that in any one direction the atom flux would be much weaker than the flux from the condensate. We estimate it to be at most 5.5% of the measured peak height for the condensate fraction. The calculated angular distribution of quantum evaporation from the non-condensate fraction in the bulk liquid is shown in Fig. 1 as a broken line. In this calculation we have used the $n(k)$ shown in the inset to find the distribution of $\mathbf{p}_{i,\parallel}$. We have assumed (1) that the quantum efficiency of quantum evaporation is the same as for the condensate and does not depend on angle or perpendicular component of momentum, and (2) the solid angle subtended by the bolometer to the point of evaporation is the same as in the experiment.

We can compare evaporation from pure ^4He with evaporation from liquid ^4He which has a fraction of a monolayer of ^3He on its free surface. With the same high-frequency phonons injected into the helium at an ambient temperature of ~ 50 mK, both ^3He and ^4He atoms are evaporated¹¹. The times of flight are in agreement with the conservation of energy expressed in equation (1), with the binding energy for the ^3He at its established value of 5.0 K (ref. 25).

The two-dimensional ^3He layer forms a condensed Fermi system with a well defined Fermi momentum and energy. ^3He atoms can be evaporated from within this Fermi disk and so there is a small energy to be subtracted from the binding energy values of 5.0 K which only applies for ^3He at the centre of the disk. But the momentum of the ^3He atoms which is parallel to the surface can be considerable, indeed greater than the parallel component of momentum of the incident phonon. Recent results (J. Warren and C. D. H. Williams, manuscript in preparation) show that the ^3He atoms are evaporated into a much larger angular range than the ^4He atoms. This indicates that the momentum of the ^3He atoms on the liquid is contributing to the parallel momentum of the evaporated atoms, which is in contrast to the behaviour of the pure ^4He discussed above.

The atoms that are quantum evaporated must come from the surface of the liquid ^4He otherwise they would be scattered by other atoms and we would not observe the narrow angular distribution shown in Fig. 1. It has been pointed out¹⁷ that the condensate fraction increases to unity as the density drops from the bulk-liquid value to zero over the $\sim 10 \text{ \AA}$ 'width' of the surface. This suggests that there should be an enhanced contribution from atoms with $\mathbf{p}_{i,\parallel} = 0$ in quantum evaporation. A search for quantum evaporation from the non-condensate fraction would help determine from which part of the density profile the atoms are evaporated, because if atoms only come from the outer parts of the surface density profile then the above-mentioned theory suggests that there is little non-condensate evaporation. $\square$

Received 16 June; accepted 24 October 1997.

1. Einstein, A. *Sitzungsber. K. Preuss. Akad. Wiss.* 3–14 (1925).
2. Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E. & Cornell, E. A. Observation of Bose-Einstein condensation in a dilute atomic vapour. *Science* **269**, 198–201 (1995).
3. Davis, K. B. *et al.* Bose-Einstein condensation in a gas of sodium atoms. *Phys. Rev. Lett.* **75**, 3969–3973 (1995).
4. Bradley, C. C., Sackett, C. A., Tollett, J. J. & Hulet, R. G. Bose-Einstein condensation of lithium: observation of limited condensate number. *Phys. Rev. Lett.* **78**, 985–989 (1997).
5. London, F. The λ -phenomenon of liquid helium and the Bose-Einstein degeneracy. *Nature* **141**, 643–644 (1938).
6. Sokol, P. E. in *Bose-Einstein Condensation* (eds Griffin, A., Snoke, D. W. & Stringari, S.) 51–85 (Cambridge, New York, 1995).
7. Bogoliubov, N. On the theory of superfluidity. *J. Phys. USSR* **11**, 23–32 (1947).
8. Campbell, C. E. in *Excitations in Two-Dimensional and Three-Dimensional Quantum Fluids* (eds Wyatt, A. F. G. & Lauter, H. J.) 159 (Plenum, New York, 1991).
9. Griffin, A. *Excitations in a Bose-Condensed Liquid* (Cambridge, London, 1993).
10. Brown, M. & Wyatt, A. F. G. The surface boundary conditions for quantum evaporation in ^4He . *J. Phys.: Condens. Matter* **2**, 5025–5046 (1990).
11. Baird, M. J., Hope, F. R. & Wyatt, A. F. G. Quantised evaporation from liquid helium. *Nature* **304**, 325–326 (1983).
12. Hope, F. R., Baird, M. J. & Wyatt, A. F. G. Quantum evaporation from liquid ^4He by rotons. *Phys. Rev. Lett.* **52**, 1528–1531 (1984).

13. Penrose, O. & Onsager, L. Bose Einstein condensation and liquid helium. *Phys. Rev.* **104**, 576–584 (1956).
14. Wyatt, A. F. G. Liquid ^4He : An ordinary and exotic liquid. *J. Phys.: Condens. Matter* **8**, 9249–9262 (1996).
15. Krotscheck, E. Liquid helium on a surface: ground state, excitations, condensate fraction and impurity potential. *Phys. Rev. B* **32**, 5713–5730 (1985).
16. Campbell, C. E. Remnants of Bose condensation and off-diagonal long range order in finite systems. *J. Low Temp. Phys.* **93**, 907–919 (1993).
17. Griffin, A. & Stringari, S. Surface region of superfluid helium as an inhomogeneous Bose-condensed gas. *Phys. Rev. Lett.* **76**, 259–263 (1996).
18. Sears, V. F., Svensson, E. C., Martel, P. & Woods, A. D. B. Neutron-scattering determination of the momentum distribution and condensate fraction in liquid ^4He . *Phys. Rev. Lett.* **49**, 279–282 (1982).
19. Sears, V. F. Kinetic energy and condensate fraction of superfluid ^4He . *Phys. Rev. B* **28**, 5109–5121 (1983).
20. Ceperley, D. M. & Pollock, E. L. Path integral computation of the low-temperature properties of liquid ^4He . *Phys. Rev. Lett.* **56**, 351–354 (1986).
21. Manousakis, E., Pandharipande, V. R. & Usmani, Q. N. Condensate fraction and momentum distribution of the ground state of liquid ^4He . *Phys. Rev. B* **31**, 7022–7028 (1985).
22. Whitlock, P. A. & Panoff, R. M. Accurate momentum distributions from computations on ^3He and ^4He . *Can. J. Phys.* **65**, 1409–1415 (1987).
23. Tucker, M. A. H. & Wyatt, A. F. G. Phonons in liquid ^4He from a heated metal film: I. The creation of high-frequency phonons. *J. Phys.: Condens. Matter* **6**, 2813–2824 (1994).
24. Tucker, M. A. H. & Wyatt, A. F. G. Phonons in liquid ^4He from a heated metal film: II The angular distribution. *J. Phys.: Condens. Matter* **6**, 2825–2834 (1994).
25. Edwards, D. O. & Saam, W. F. The free surface of liquid helium. *Prog. Low Temp. Phys.* **7**, 283–369 (1978).

Acknowledgements. I thank A. Griffin for encouragement and for comments on the draft manuscript; C. Williams and J. Warren for discussions on their results for ^3He and allowing me to quote them; C. Williams and M. Brown for the simulation of the peak in Fig. 1; and M. Gibbs for discussions on neutron scattering. This work was supported by the EPSRC.

Correspondence should be addressed to the author (e-mail: A.F.G.Wyatt@exeter.ac.uk).

Electronic structure of atomically resolved carbon nanotubes

Jeroen W. G. Wildöer*, Liesbeth C. Venema*, Andrew G. Rinzler†, Richard E. Smalley† & Cees Dekker*

* Department of Applied Physics and DIMES, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, The Netherlands

† Center for Nanoscale Science and Technology, Rice Quantum Institute, Departments of Chemistry and Physics, MS-100, Rice University, PO Box 1892, Houston, Texas 77251, USA

Carbon nanotubes can be thought of as graphitic sheets with a hexagonal lattice that have been wrapped up into a seamless cylinder. Since their discovery in 1991¹, the peculiar electronic properties of these structures have attracted much attention. Their electronic conductivity, for example, has been predicted^{2–4} to depend sensitively on tube diameter and wrapping angle (a measure of the helicity of the tube lattice), with only slight differences in these parameters causing a shift from a metallic to a semiconducting state. In other words, similarly shaped molecules consisting of only one element (carbon) may have very different electronic behaviour. Although the electronic properties of multi-walled and single-walled nanotubes^{5–12} have been probed experimentally, it has not yet been possible to relate these observations to the corresponding structure. Here we present the results of scanning tunnelling microscopy and spectroscopy on individual single-walled nanotubes from which atomically resolved images allow us to examine electronic properties as a function of tube diameter and wrapping angle. We observe both metallic and semiconducting carbon nanotubes and find that the electronic properties indeed depend sensitively on the wrapping angle. The bandgaps of both tube types are consistent with theoretical predictions. We also observe van Hove singularities at the onset of one-dimensional energy bands, confirming the strongly one-dimensional nature of conduction within nanotubes.

As shown in Fig. 1A, a carbon nanotube can be constructed by wrapping up a single sheet of graphite such that two equivalent sites

of the hexagonal lattice coincide. The wrapping vector $\mathbf{C}$, which defines the relative location of the two sites, is specified by a pair of integers (n, m) that relate $\mathbf{C}$ to the two unit vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ ($\mathbf{C} = n\mathbf{a}_1 + m\mathbf{a}_2$). A tube is called 'armchair' if n equals m , and 'zigzag' in the case $m = 0$. All other tubes are of the 'chiral' type and have a finite wrapping angle ϕ with $0^\circ < \phi < 30^\circ$ (ref. 13). To be

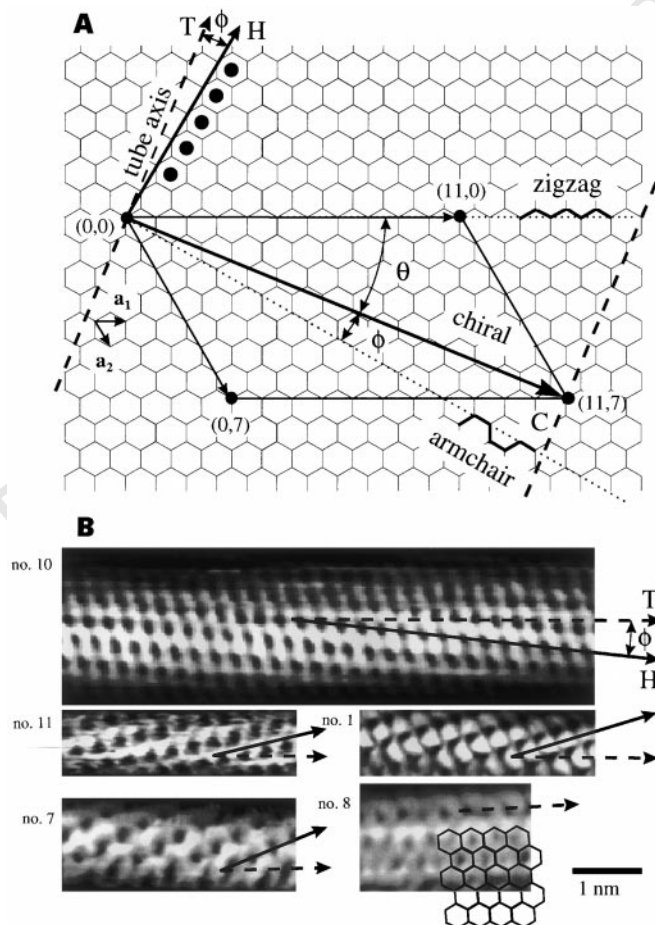


Figure 1 Relation between the hexagonal carbon lattice and the chirality of carbon nanotubes. **A**, the construction of a carbon nanotube from a single graphene sheet. By rolling up the sheet along the wrapping vector $\mathbf{C}$, that is, such that the origin $(0,0)$ coincides with point $\mathbf{C}$, a nanotube indicated by indices $(11,7)$ is formed. Wrapping vectors along the dotted lines lead to tubes that are zigzag or armchair. All other wrapping angles lead to chiral tubes whose wrapping angle is specified relative to either the zigzag direction (θ) or to the armchair direction ($\phi = 30^\circ - \theta$). Dashed lines are perpendicular to $\mathbf{C}$ and run in the direction of the tube axis indicated by vector $\mathbf{T}$. The solid vector $\mathbf{H}$ is perpendicular to the armchair direction and specifies the direction of nearest-neighbour hexagon rows indicated by the black dots. The angle between $\mathbf{T}$ and $\mathbf{H}$ is the chiral angle ϕ . **B**, Atomically resolved STM images of individual single-walled carbon nanotubes. The lattice on the surface of the cylinders allows a clear identification of the tube chirality. Dashed arrows represent the tube axis $\mathbf{T}$ and the solid arrows indicate the direction of nearest-neighbour hexagon rows $\mathbf{H}$. Tubes no. 10, 11 and 1 are chiral, whereas tubes no. 7 and 8 have a zigzag and armchair structure, respectively. Tube no. 10 has a chiral angle $\phi = 7^\circ$ and a diameter $d = 1.3$ nm, which corresponds to the $(11,7)$ type of panel **A**. A hexagonal lattice is plotted on top of image no. 8 to clarify the non-chiral armchair structure. Carbon nanotubes were synthesized as described in ref. 14. TEM studies¹⁴ have shown that the material consists mainly of ~ 1.4 -nm-thick single-walled nanotubes. These were deposited from a dispersion in 1,2 dichloroethane on single-crystalline Au(111) facets. Topographic images were obtained by recording the tip height at constant tunnel current in a home-built STM²⁵ operated at 4 K. The Pt/Ir tips were cut in ambient air by scissors. Typical bias parameters are those of image no. 10, that is, a tunnel current $I = 60$ pA, and a bias voltage $V_{\text{bias}} = 500$ mV.

able to correlate the electronic properties of a tube with its atomic structure, atomically resolved images are required. In Fig. 1B we show examples of such scanning tunnelling microscopy (STM) images of single-wall carbon nanotubes. (The chiral angle $\phi = 7^\circ$, and diameter $d = 1.3$ nm of tube no. 10 correspond to vector $C = (11, 7)$ in the top panel of Fig. 1.)

The critical dependence of the electronic spectra of nanotubes on the tube indices (n, m) can be understood by again taking the two-dimensional graphene sheet as a starting point. In the circumferential direction (along C), periodic boundary conditions apply, that is, $C \cdot k = 2\pi q$ (ref. 13), where k is the wavevector and q is an integer. This leads to a set of allowed values for k which can be substituted into the energy dispersion relation for a graphene sheet to yield the dispersion relations for the tube, with q labelling the various one-dimensional modes. Calculations¹³ predict that armchair ($n = m$) tubes have bands crossing the Fermi level and are therefore metallic. For all other tubes (chiral and zigzag) there exist two possibilities. When $n - m = 3l$ (where l is an integer), tubes are also expected to

be metallic. In the case $n - m \neq 3l$, tubes are predicted to be semiconducting with an energy gap of the order of ~ 0.5 eV. This gap should only depend on the diameter, that is, $E_{\text{gap}} = 2\gamma_0 a_{C-C}/d$, where γ_0 is the C–C tight-binding overlap energy, a_{C-C} the nearest-neighbour C–C distance (0.142 nm) and d the diameter.

Atomic resolution was achieved on more than 20 tubes, corresponding to $\sim 80\%$ of the tubes investigated. In most cases only a few rows of atoms on top of the tube could be imaged. Very few armchair tubes were observed, which is in apparent contrast with earlier investigations on the same type of material^{14,15}. However, unlike these earlier experiments we have concentrated on individual tubes, and have neglected the ropes of tubes.

Figure 1B shows atomically resolved STM images on five different tubes. The most prominent feature is the triangular lattice of dark dots with a lattice spacing of ~ 0.25 nm, as expected for a graphene lattice. We attribute the dark dots to the centres of the hexagons¹⁶. The atomic resolution images allow us unambiguously to determine the chiral angle, that is, the angle between the hexagon rows and the

Table 1 Overview of the experimental STM and STS results on individual single-walled carbon nanotubes

Tube number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Chiral angle ϕ (deg)	25	4	7	24	9	14	30	0		7	14		7	16		4	9	16	6	29	16	18	9	7	28	27	6
Diameter d (nm)	1.4	1.4	2.0	1.2	1.7	1.3	1.1	1.3	1.3	1.3	1.3	1.5	1.5	1.4	1.4	1.4	1.4	1.2	1.3	1.3	1.9	1.0	1.4	1.5	1.7	1.4	1.4
E_{gap} (eV)	0.55	0.60	0.50	0.65	1.7	1.8	1.9		0.65			1.8	2.0	1.9	0.60	0.5	0.55	0.60	0.60	0.5	0.4						
δE (eV)	0.25	0.3	0.25	0.3	0.3	0.2	0.2		0.30			0.3	0.4	0.3	0.30	0.2	0.30	0.30	0.30	0.2	–0.2*						

Here d is the nanotube diameter; ϕ is the chiral angle; E_{gap} is the apparent bandgap in the STS- V spectra and δE is the shift of the Fermi energy due to doping of the tube by the substrate. Note that a chiral angle of 0° denotes an armchair nanotube, and an angle of 30° a zigzag tube. The flat Au surface allowed the diameter d of the nanotubes to be determined with an accuracy of 0.1 nm by measuring the tube heights relative to the surface. A possible systematic uncertainty in determining the diameter is due to a difference in barrier heights for the gold substrate and the tubes. The wrapping angle ϕ can be determined with an accuracy of $\sim 1^\circ$. Accuracy in ϕ is limited by the curvature of the tubes. A combination of high accuracy in both ϕ ($\sim 1^\circ$) and d (~ 0.05 nm) is required for an unambiguous identification of the n, m indices. Accuracy in E_{gap} and δE is 0.05–0.1 eV.

* For this sample we observed a shift in the Fermi energy towards the conduction band, instead of a shift towards the valence band as observed in the other samples. We speculate that the gold substrate for this sample may have had an anomalously low work function.

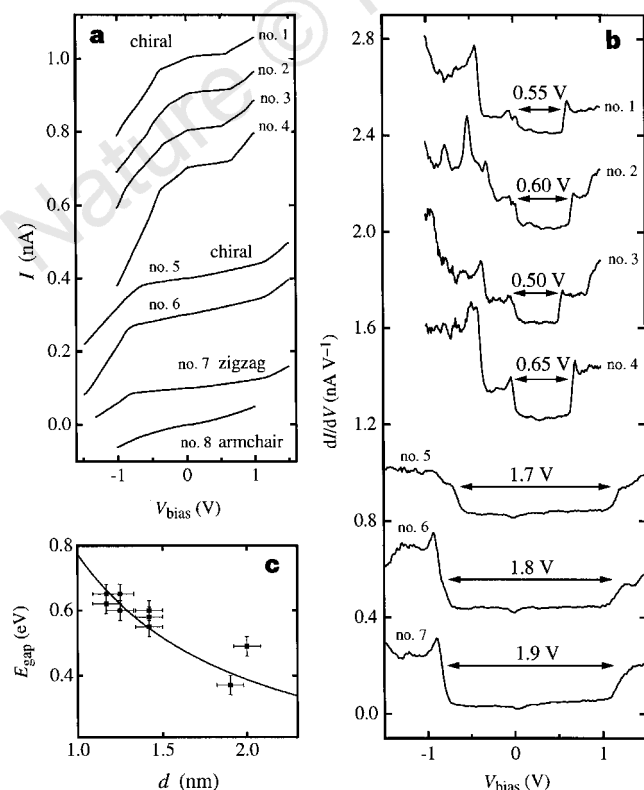


Figure 2 Electronic properties of single-walled carbon nanotubes. **a**, Current-voltage curves obtained by tunnelling spectroscopy on various individual nanotubes. Tubes nos 1–6 are chiral, no. 7 is zigzag and no. 8 is armchair. The bias voltage is applied to the sample, which means that the sign of V_{bias} corresponds to that of the energy relative to the tube Fermi level. Curves nos 1–7 show a low conductance at low bias, followed by several kinks at larger bias voltages. The armchair tube does not show clear kinks in the range -1 to $+1$ V. **b**, The derivatives dI/dV . For clarity, curves are offset vertically by multiples of 0.4 nA V^{-1} . Gaps are indicated by the arrows. Two categories can be distinguished: one with gap values around 0.5–0.6 eV, the other with significantly larger gap values. The first category of tubes is identified as the semiconducting type, the second as metallic tubes. About 12 out of 18 tubes were semiconducting, in accordance with the expected ratio of 2 out of 3. We note that these tubes, besides the primary gaps, also show peaks associated with secondary and higher-order gaps. For the secondary gaps we find: 1.4 eV (no. 1), 1.1 eV (no. 2), 1.2 eV (no. 3), 1.4 eV (no. 4). All gaps seem to have shifted to the right, which indicates doping of the tubes by the substrate. We refrain from concluding a zero or finite DOS from the value of dI/dV in the gap. Possible mechanisms for a small finite dI/dV within the semiconducting gap are, for example, tip-induced band bending or residual tunnelling through the tube to the gold substrate. The doping behaviour, however, indicates a finite DOS for metallic tubes, and a zero DOS for semiconducting tubes. The small dips in dI/dV at zero bias present in some of the curves are not yet understood. The displayed dI/dV data are the result of averaging over ~ 50 individual I - V curves for improved signal-to-noise ratio. The individual curves all contain the same essential features. **c**, Energy gap E_{gap} versus diameter d for semiconducting chiral tubes. The data points correspond quite well to the theoretical predicted values. The solid line denotes a fit of $E_{\text{gap}} = 2\gamma_0 a_{C-C}/d$ with $\gamma_0 = 2.7$ eV. Tunnel currents I as a function of the bias voltage V applied to the sample were recorded with a home-built STM (ref. 25) while scanning and feedback were switched off. The Pt/Ir tips were cut in ambient air by scissors.

tube axis. This enables us to distinguish chiral tubes (such as shown in images nos 1, 10 and 11 in Fig. 1B) from zigzag (no. 7) and armchair (no. 8) tubes. A wide variety of chiral angles is observed (see Table 1), in contrast to earlier electron-diffraction studies on ropes of single-walled nanotubes¹⁵.

The vacuum barrier between the STM tip and the sample forms a convenient junction for scanning tunnelling spectroscopy (STS) as it allows tunnel currents at large bias voltages. In STS, scanning and feedback are switched off, and current I is recorded as a function of the bias voltage V applied to the sample. The differential conductance (dI/dV) can then be considered to be proportional to the density of states (DOS) of the tube examined. Before and after taking STS measurements on a tube, reference measurements were performed on the gold substrate. Only when the curves on gold were approximately linear, and did not show kinks or steps, were data on a tube recorded. I - V traces were only taken far from the ends of the tube and when tubes were isolated from each other. On all the tubes reported here, STS curves taken at different positions (typically over ~ 40 nm) showed consistent features.

Figure 2a shows a selection of I - V curves obtained by STS on different tubes. Most curves show a low conductance at low bias, followed by several kinks at larger bias voltages. From all the chiral tubes that we have investigated, we can clearly distinguish two categories: the one has a well defined gap value around 0.5–0.6 eV, the other has significantly larger gap values of ~ 1.7 –2.0 eV (see Table 1 and Fig. 2b). The gap values of the first category coincide very well with the expected gap values for semiconducting tubes. As illustrated in Fig. 2c, which displays gap versus tube diameter, the measurement agrees well with theoretical gap values obtained for an overlap energy $\gamma_0 = 2.7 \pm 0.1$ eV, which is close to the value $\gamma_0 = 2.5$ eV suggested for a single graphene sheet^{13,17,18}. The very large gaps that we observe for the second category of tubes, 1.7–2.0 eV, are in good agreement with the values 1.6–1.9 eV that we obtain from one-dimensional dispersion relations¹³ for a number of

metallic tubes ($n - m = 3l$) of ~ 1.4 nm diameter. Metallic nanotubes are expected to have a small but finite DOS near the Fermi energy (E_F) and the apparent 'gap' is associated with DOS peaks at the band edges of the next one-dimensional modes.

The gaps seen in Fig. 2 are not symmetrically positioned around zero bias voltage. This shows that the tubes are doped by charge transfer from the Au(111) substrate. The latter has a work function of ~ 5.3 eV which is much higher than that of the nanotubes (which presumably is similar to the 4.5-eV work function of graphite). This shifts the Fermi energy towards the valence band of the tube. In the semiconducting tubes, the Fermi energy seems to have shifted from the centre of the gap to the valence band edge. In the metallic tubes it is shifted by ~ 0.3 eV, which is much lower than half of the 'gap'. This provides experimental evidence that these tubes indeed have a finite DOS within the 'gap', in contrast to the zero DOS for semiconducting tubes.

The chiral tubes thus seem to be either semiconducting or metallic, with gap values as predicted. The data provide a striking verification of the band-structure calculations of nanotube electronic properties. Our electronic spectra for armchair and zigzag tubes are also consistent with the calculations, but the small number of such tubes in our experiments prevent a more general statement. For chiral metallic tubes, it has been suggested¹⁹ that a small (~ 0.01 eV) gap will open up owing to the tube curvature. We did not observe such a small gap at the centre of the large 'gap' of chiral metallic tubes. This may be attributed to hybridization between wavefunctions of the tube and the gold substrate²⁰ which causes smoothening of small-energy features.

Sharp Van Hove singularities in the DOS are predicted at the onsets of the subsequent energy bands, reflecting the one-dimensional character of carbon nanotubes. The derivative spectra indeed show a number of peak structures (Fig. 2b). The peaks we observe differ in height depending on the configuration of the STM tip. On semiconductors, $(dI/dV)/(I/V)$ has been argued to give a better representation of the DOS than the direct derivative dI/dV , partly because the normalization accounts for the voltage dependence of the tunnel barrier at high bias^{21–24}. In Fig. 3 we show $(dI/dV)/(I/V)$ on the ordinate. Sharp peaks are observed with a shape that resembles that predicted for Van Hove singularities (see right inset of Fig. 3): with increasing $|V|$, $(dI/dV)/(I/V)$ rises steeply, followed by a slow decrease. The latter should be $\propto 1/\sqrt{|V|}$ according to the theory for the DOS near a one-dimensional band edge. The experimental peaks have a finite height and are broadened, which again can be attributed to hybridization of wave functions.

Our results constitute, to the best of our knowledge, the first experimental test of the vast amount of band-structure calculations that have appeared in recent years. The central theoretical prediction, that chiral tubes are either semiconducting or metallic depending on minor variations of wrapping angle or diameter, has been verified. □

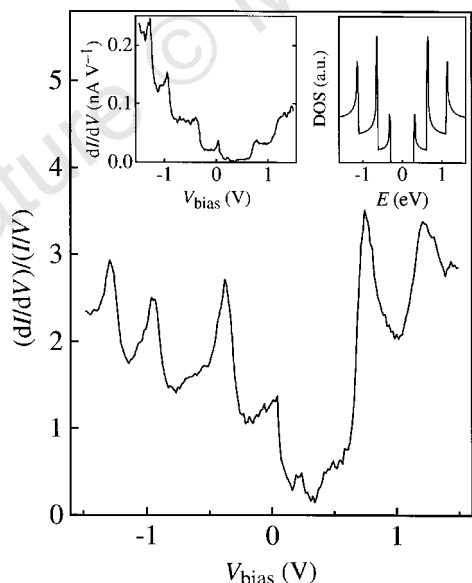


Figure 3 $(dI/dV)/(I/V)$ which is a measure of the density of states versus V for nanotube no. 9. The asymmetric peaks correspond to Van Hove singularities at the onsets of one-dimensional energy bands of the carbon nanotube. The left inset displays the raw dI/dV data. The right inset is the calculated density of states (DOS) for a (16,0) tube, which displays a typical example of the peak-like DOS for a semiconducting tube (a.u., arbitrary units). The experimental peaks have a finite height and are broadened, which we attribute to hybridization between the wavefunctions of the tube and the gold substrate. The overall shape of the experimental peaks however still resembles that predicted by theory.

Received 3 September; accepted 29 October 1997.

1. Iijima, S. Helical microtubules of graphitic carbon. *Nature* **354**, 56–58 (1991).
2. Hamada, N., Sawada, S.-I. & Oshiyama, A. New one-dimensional conductors: graphite microtubules. *Phys. Rev. Lett.* **68**, 1579–1581 (1992).
3. Saito, R., Fujita, M., Dresselhaus, G. & Dresselhaus, M. S. Electronic structure of chiral graphene tubules. *Appl. Phys. Lett.* **60**, 2204–2206 (1992).
4. Mintmire, J. W., Dunlap, B. I. & White, C. T. Are fullerene tubules metallic? *Phys. Rev. Lett.* **68**, 631–634 (1992).
5. Tans, S. J. *et al.* Individual single-wall carbon nanotubes as quantum wires. *Nature* **386**, 474–477 (1997).
6. Bockrath, *et al.* Single-electron transport in ropes of carbon nanotubes. *Science* **275**, 1922–1925 (1997).
7. Ebbsen, T. W. *et al.* Electrical conductivity of individual carbon nanotubes. *Nature* **382**, 54–56 (1996).
8. Ge, M. & Sattler, K. Vapor-condensation generation of and STM analysis of fullerene tubes. *Science* **260**, 515–518 (1993).
9. Zhang, Z. & Lieber, C. M. Nanotube structure and electronic properties probed by scanning tunneling microscopy. *Appl. Phys. Lett.* **62**, 2792–2794 (1993).
10. Ge, M. & Sattler, K. STM of single-shell nanotubes of carbon. *Appl. Phys. Lett.* **65**, 2284–2286 (1994).
11. Olk, C. H. & Heremans, J. P. Scanning tunneling spectroscopy of carbon nanotubes. *J. Mater. Res.* **9**, 259–262 (1994).

12. Carroll, D. L. *et al.* Electronic structure and localized states at carbon nanotube tips. *Phys. Rev. Lett.* **78**, 2811–2814 (1997).
13. Dresselhaus, M. S., Dresselhaus, G. & Eklund, P. C. *Science of Fullerenes and Carbon Nanotubes* (Academic, San Diego, 1996).
14. Thess, A. *et al.* Crystalline ropes of metallic carbon nanotubes. *Science* **273**, 483–487 (1996).
15. Conley, J. M., Nikolaev, P., Thess, A. & Smalley, R. E. Electron nano-diffraction study of carbon single-walled nanotube ropes. *Chem. Phys. Lett.* **265**, 379–384 (1997).
16. Venema, L. C. *et al.* STM atomic resolution images of single-wall carbon nanotubes. *Appl. Phys. A* (in the press).
17. Mintmire, J. W., Robertson, D. H. & White, C. T. Properties of fullerene nanotubes. *J. Phys. Chem. Solids* **54**, 1835–1840 (1993).
18. White, C. T. *et al.* in *Buckminsterfullerenes* (eds Billups, W. E. & Ciufolini, M. A.) 125–184 (VCH, Weinheim, 1993).
19. Kane, C. L. & Mele, E. J. Size, shape, and low energy electronic structure of carbon nanotubes. *Phys. Rev. Lett.* **78**, 1932–1936 (1997).
20. Gadzuk, J. W. Resonance-tunneling spectroscopy of atoms adsorbed on metal surfaces: theory. *Phys. Rev. B* **1**, 2110–2129 (1970).
21. Strosio, J. A., Feenstra, R. M. & Fein, A. P. Electronic structure of the Si(111)2 × 1 surface by scanning tunneling spectroscopy. *Phys. Rev. Lett.* **57**, 2579–2582 (1986).
22. Feenstra, R. M., Strosio, J. A. & Fein, A. P. Tunneling spectroscopy of the Si(111)2 × 1 surface. *Surf. Sci.* **181**, 295–306 (1987).
23. Lang, N. D. Spectroscopy of single atoms in the scanning tunneling microscope. *Phys. Rev. B* **34**, R5947–R5950 (1986).
24. Hamers, R. J. in *Scanning Tunneling Microscopy and Spectroscopy* (ed. Bonnell, D. A.) 51–103 (VCH, New York, 1993).
25. Wildöer, J. W. G., van Roij, A. J. A., van Kempen, H. & Harmans, C. J. P. M. Low-temperature scanning tunneling microscope for use on artificially fabricated nanostructures. *Rev. Sci. Instrum.* **65**, 2849–2852 (1994).

Acknowledgements. The first two authors contributed equally to the present work. We thank S. J. Tans and J. C. Charlier for discussions, and L. P. Kouwenhoven, J. E. Mooij and P. M. Dewilde for support. The work at Delft was supported by the Dutch Foundation for Fundamental Research of Matter (FOM). The nanotube research at Rice was funded in part by the US NSF, the Texas Advanced Technology Program and the Robert A. Welch Foundation.

Correspondence and requests for materials should be addressed to C.D. (e-mail: dekker@qt.tn.tudelft.nl).

Atomic structure and electronic properties of single-walled carbon nanotubes

Teri Wang Odom*, Jin-Lin Huang*, Philip Kim† & Charles M. Lieber*†

* Department of Chemistry and Chemical Biology, and † Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

Carbon nanotubes¹ are predicted to be metallic or semiconducting depending on their diameter and the helicity of the arrangement of graphitic rings in their walls^{2–5}. Scanning tunnelling microscopy (STM) offers the potential to probe this prediction, as it can resolve simultaneously both atomic structure and the electronic density of states. Previous STM studies of multi-walled nanotubes^{6–9} and single-walled nanotubes (SWNTs)¹⁰ have provided indications of differing structures and diameter-dependent electronic properties, but have not revealed any explicit relationship between structure and electronic properties. Here we report STM measurements of the atomic structure and electronic properties of SWNTs. We are able to resolve the hexagonal-ring structure of the walls, and show that the electronic properties do indeed depend on diameter and helicity. We find that the SWNT samples exhibit many different structures, with no one species dominating.

The diameter and helicity of a defect-free SWNT are uniquely characterized by the vector $\mathbf{c}_h = n\mathbf{a}_1 + m\mathbf{a}_2 \equiv (n, m)$ that connects crystallographically equivalent sites on a two-dimensional graphene sheet, where $\mathbf{a}_1$ and $\mathbf{a}_2$ are the graphene lattice vectors and n and m are integers (Fig. 1). Electronic band structure calculations^{2–5} predict that the (n, m) indices determine the metallic or semiconducting behaviour of SWNTs. Zigzag $(n, 0)$ SWNTs should have two distinct types of behaviour: the tubes will be metals when $n/3$ is an integer, and otherwise semiconductors^{3–5}. As $\mathbf{c}_h$ rotates away from

$(n, 0)$, chiral (n, m) SWNTs are possible with electronic properties similar to the zigzag tubes; that is, when $(2n + m)/3$ is an integer the tubes are metallic, and otherwise semiconducting. The gaps of the semiconducting $(n, 0)$ and (n, m) tubes should depend inversely on diameter. Finally, when $\mathbf{c}_h$ rotates 30° relative to $(n, 0)$, $n = m$. The (n, n) or armchair tubes are expected to be truly metallic with band crossings at $\mathbf{k} = \pm 2/3$ of the one-dimensional Brillouin zone. It has been suggested that SWNT samples produced by laser vaporization¹¹ and arc¹² methods consist predominantly of $(10, 10)$ metallic armchair tubes.

We have carried out STM measurements in ultra-high vacuum at 77 K on purified SWNT samples produced by laser vaporization¹¹. Typical atomically resolved images of a SWNT on the surface of a rope, which consists of parallel tubes¹¹, and isolated SWNTs on a Au(111) substrate are shown in Fig. 2a and b, respectively. Figure 2a shows the expected honeycomb lattice for a SWNT with a C–C spacing of 0.14 ± 0.02 nm. The chiral angle is readily determined by identifying the zigzag tube axis direction (the line connecting sites separated by 0.426 nm) relative to the sample tube axis. This shows quite clearly that the tube is chiral with an axis orientated at an angle of $-8.0 \pm 0.5^\circ$ relative to that for a zigzag nanotube. As the tube axis is perpendicular to $\mathbf{c}_h$, this corresponds to the angle between $\mathbf{c}_h$ and $(n, 0)$ in Fig. 1. From this angle and the measured diameter of 1.0 ± 0.05 nm, we can assign (n, m) indices of either $(11, 2)$ or $(12, 2)$; the angle/diameter for $(11, 2)$ and $(12, 2)$ are $-8.2^\circ/0.95$ nm and $-7.6^\circ/1.03$ nm, respectively. We note that an $(11, 2)$ tube is expected to be metallic, whereas a $(12, 2)$ tube should be semiconducting. The helicity of the lower isolated SWNT in Fig. 2b was determined in a similar manner, yielding a chiral angle of $-11.0 \pm 0.5^\circ$; the diameter of this tube is 1.08 ± 0.05 nm. These parameters match closely the values expected for a $(12, 3)$ tube, $\pm 10.9^\circ/1.08$ nm, and reasonably exclude other choices of indices.

Central to the work reported here is our ability to characterize the electronic properties of the atomically resolved nanotubes by tunnelling spectroscopy. Specifically, current (I) versus voltage (V) was measured at specific sites along the tubes and differentiated to yield the normalized conductance, $(V/I)dI/dV$, which has been shown¹³ to provide a good measure of the main features in the local density of electronic states (LDOS) for metals and semiconductors. The gradual increase in current in the I – V data (Fig. 2c, d) recorded on the SWNTs imaged in Fig. 2a, b shows qualitatively that both tubes are metallic. The LDOS determined from data sets recorded at different locations along the tubes are very similar, demonstrating the reproducibility of the measurements; furthermore, the LDOS for both tubes are roughly constant between -600 and $+600$ mV as expected for a metal. Small variations in the LDOS with energy are not significant and arise from noise in the data. These spectroscopy results are similar to those obtained on the Au(111) substrate except that the surface state 450 meV below the Fermi level¹⁴ is also observed on the latter.

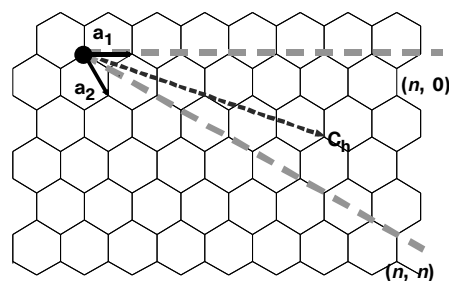


Figure 1 Schematic of a two-dimensional graphene sheet illustrating lattice vectors $\mathbf{a}_1$ and $\mathbf{a}_2$, and the roll-up vector $\mathbf{c}_h = n\mathbf{a}_1 + m\mathbf{a}_2$. The limiting cases of $(n, 0)$ zigzag and (n, n) armchair tubes are indicated with dashed lines. As represented here, the angle between the zigzag configuration and $\mathbf{c}_h$ is negative.

The metallic behaviour of our (12,3) tube is in agreement with the prediction that $(2n + m)/3$ is an integer, and additionally suggests that the indices for the tube in Fig. 2a are (11,2) rather than (12,2). We have also characterized a metallic, achiral zigzag SWNT with a diameter of 0.95 ± 0.05 nm. This diameter is very close to the expected 0.94 nm diameter of a (12,0) tube, although possibly indistinguishable from the 1.02 nm diameter expected for a (13,0) tube. There are two other important points that these data address. First, curvature in the graphene sheet of a SWNT should cause the π/σ bonding and π^*/σ^* antibonding orbitals on carbon to mix and create a small gap at the Fermi level in these metallic tubes^{3,5}. We have not observed evidence for this small gap, although it is possible that the thermal energy at 77 K, 7 meV, smears the gap structure predicted to be of the order of 8 meV for a (12,0) tube³. Second, the LDOS recorded on metallic SWNTs in a rope and isolated on the substrate are similar, thus suggesting that inter-tube interactions do not perturb the electronic structure on an energy scale of 77 K.

We have also characterized a number of semiconducting SWNTs in our studies. Indeed, more than half of the SWNTs observed either as isolated tubes or in ropes were found to be moderate gap semiconductors. A typical example of the atomically resolved structural and tunnelling spectroscopy data obtained from isolated SWNTs is shown in Fig. 3. Analysis of the image (Fig. 3a) shows that the upper tube has a chiral angle of $11.2 \pm 0.5^\circ$ (that is, opposite helicity to the tubes in Fig. 2) and a diameter of 0.95 ± 0.05 nm.

These angle/diameter constraints agree best with the $11.7^\circ/1.0$ nm for a (14,-3) tube, although the $10.9^\circ/1.08$ nm angle/diameter of the next closest (15,-3) indices are close to our uncertainty. The $I-V$ data recorded with this atomic-resolution image (Fig. 3b inset) shows distinctly different behaviour from the metallic tubes and is consistent with a semiconductor; that is, the current is very small for $-300 \leq V \leq +400$ mV but increases sharply when $|V|$ is increased further. The calculated $(V/I)dI/dV$ shows sharp increases at -325 and +425 mV that correspond to the conduction and valence band edges in the LDOS, and thus we assign a bandgap of 750 meV.

The observed semiconducting behaviour is consistent with the expectation that a (14,-3) tube should be a moderate gap semiconductor (that is, $(2n + m)/3$ is not an integer). In addition, we have observed similar semiconducting behaviour for other chiral and zigzag tubes characterized with atomic resolution. A summary of the energy gaps (E_g) obtained from these measurements for tubes with diameters between 0.6 and 1.1 nm is shown in Fig. 3c. These results show the expected¹ $1/\text{diameter}$ (d) dependence, and can be fitted to $E_g = 2\gamma_0 a_{C-C}/d$, where $\gamma_0 = 2.45$ eV is the nearest-neighbour overlap integral and a_{C-C} is the C-C distance. Significantly, this value of γ_0 is in good agreement with the value (2.5 eV) determined from calculations¹, and provides an additional consistency check in this work.

Our observation of semiconducting and metallic SWNTs with subtle changes in structure clearly confirm the remarkable electronic behaviour of the nanotubes that may be exploited in future

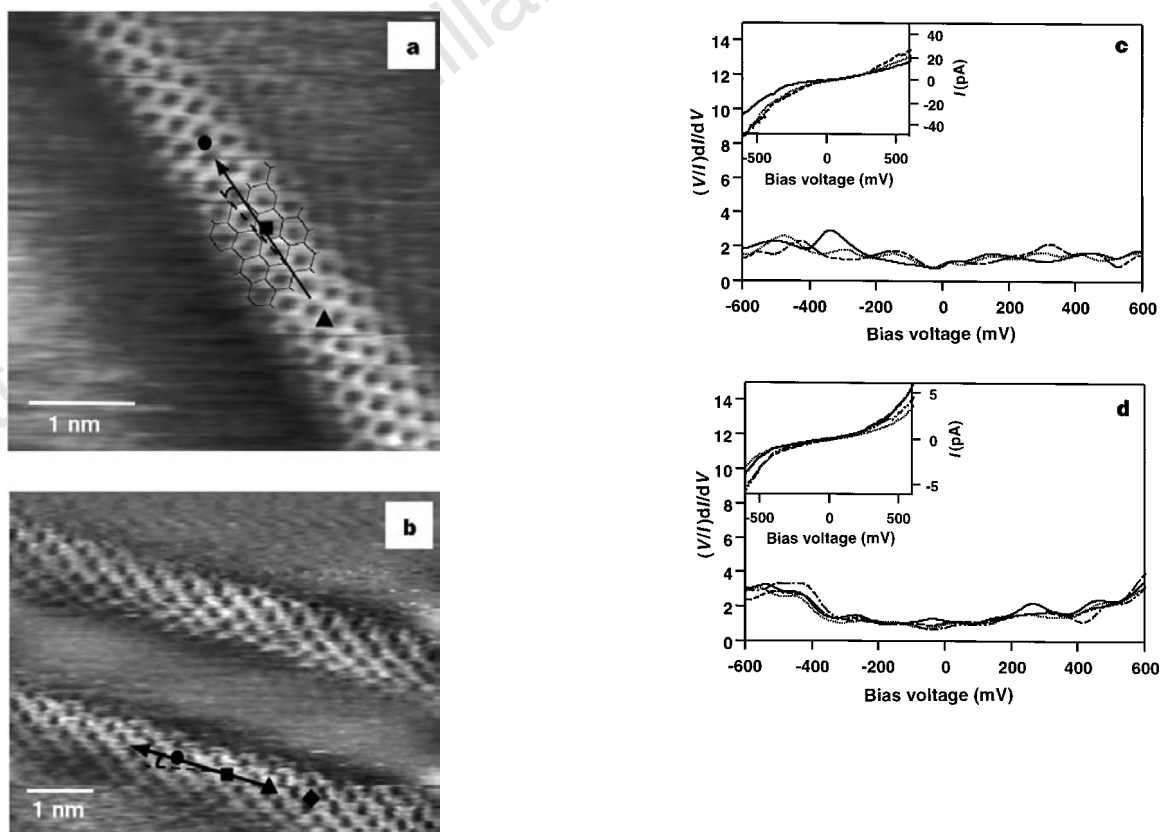


Figure 2 Atomic structure and spectroscopy of metallic SWNTs. STM images of **a**, a SWNT exposed at the surface of a rope and **b**, isolated SWNTs on a Au(111) substrate. The images were recorded in the constant-current mode with bias voltages of 50 and 150 mV, respectively, and tunnelling current of 150 pA. The images were low-pass filtered. The tube axes in both images are indicated with solid, black arrows, and the zigzag direction are highlighted by dashed lines. A portion of a two-dimensional graphene layer is overlaid in **a** to highlight the

atomic structure. The symbols in **a** and **b** correspond to the locations where $I-V$ were measured. The two parallel tubes in **b** could correspond to distinct tubes or to an image of one tube by two tips, although the interpretation of our results is not affected by either explanation. **c**, **d**, Calculated normalized conductance, $(V/I)dI/dV$, and measured $I-V$ (inset) from the locations indicated in **a**, **b**: —/●; ---/■; - - -/▲; .../◆. The feature at -400 mV in **d** was not observed for every metallic tube, and may arise from a nanotube-substrate interaction.

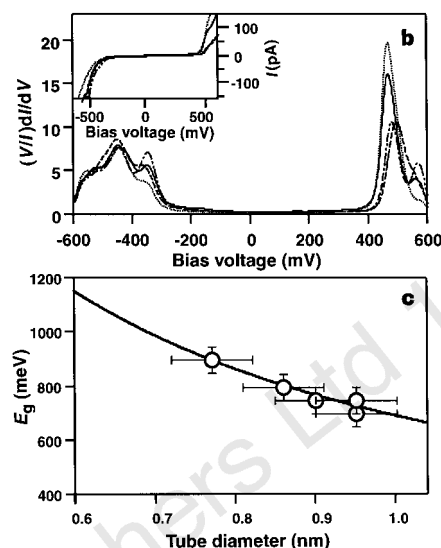
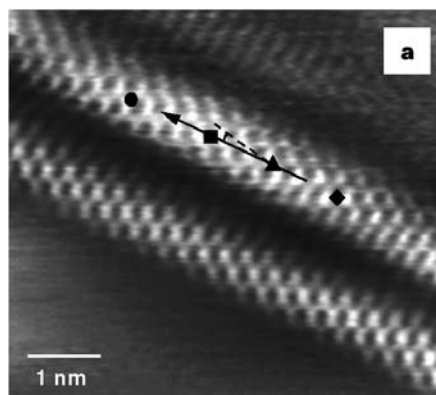


Figure 3 Structure and spectroscopy of semiconducting SWNTs. **a**, Constant current image of isolated SWNTs on a Au(111) surface recorded with a bias voltage of 300 mV and a tunnelling current of 150 pA. The solid, black arrow highlights the tube axis, and the dashed line indicates the zigzag direction. **b**, Calculated normalized conductance and measured I - V (inset) data from the

positions indicated by the symbols in **a**; the symbols are the same as in Fig. 2. **c**, Summary of energy gap (E_g) versus tube diameter data obtained in these studies. The angles/diameters of these points may be fitted with (10,0), (11,0), (11,0), (14,-3) and (13,-1). The solid line corresponds to the fit described in the text.

applications. We believe that these results also have significant implications for the present. First, the data show a richness of structures (and electronic properties), and indicate that no one SWNT type dominates. This observation contrasts with previous suggestions^{11,15} that SWNTs prepared by laser vaporization consist primarily of (10,10) armchair tubes. Because our STM measurements are biased towards SWNTs that have separated from the outer portion of ropes, it is possible that this discrepancy reflects a rope structure in which (10,10) armchair tubes lie at the core and are surrounded by structurally diverse tubes. We believe that it will be possible to test this idea by cutting and etching the ropes (J. Liu, R. E. Smalley, T.W.O., J.L.H., P.K. and C.M.L., work in progress); such results could help to elucidate the growth mechanism.

The presence of a large fraction of semiconducting nanotubes should be considered when interpreting electrical measurements that probe the bulk properties of ropes, such as the temperature-dependent resistivity¹⁶ and doping-induced changes in electrical conductivity¹⁷. For example, the cross-over in resistivity with temperature¹⁶ could have a trivial explanation in terms of variable-range hopping between metallic tubes separated by semiconducting tubes. We believe that STM characterization combined with growth and purification studies will provide a rational pathway for producing structurally homogeneous samples of nanotubes in the future. □

Methods

SWNT samples were prepared by laser vaporization¹¹, and deposited by spin-coating organic suspensions onto Au(111) surfaces. Immediately after deposition, the substrate was loaded into an ultra-high vacuum (UHV) chamber and transferred under vacuum to the UHV STM that was stabilized at 77 K. Atomic force microscopy and field-emission scanning electron microscopy showed that the samples consisted primarily of SWNT ropes spaced $\sim 10 \mu\text{m}$ apart. Imaging and spectroscopy measurements were made using etched tungsten tips with the bias (V) applied to the tip. The resolution and calibration of the STM were confirmed *in situ* by imaging the Au(111) atomic lattice and steps. The nanotube diameters were determined by deconvolution of the tip radius from the observed images; typically, 400 cross-sections were averaged and used in the deconvolution. The parameters obtained from this model were reasonable and in agreement with those obtained from Au steps in the same experiments.

Spectroscopy measurements were made by recording and averaging 5–10 I - V curves at specific locations on atomically resolved SWNTs. Typically, 6–8 distinct locations were measured for a given atomic image. To ensure the reliability of these measurements we routinely checked that clean areas of the Au(111) substrate showed typical metallic I - V behaviour and the expected two-dimensional surface state¹⁴, and that I varied exponentially with tip-sample separation. The normalized conductance $(V/I)dI/dV$ was calculated from digital I - V data using standard methods¹³.

Received 10 October; accepted 26 November 1997.

- Dresselhaus, M. S., Dresselhaus, G. & Eklund, P. C. *Science of Fullerenes and Carbon Nanotubes* (Academic, San Diego, 1996).
- Mintmire, J. W., Dunlap, B. I. & White, C. T. Are fullerene tubes metallic? *Phys. Rev. Lett.* **68**, 631–634 (1992).
- Hamada, N., Sawada, S. & Oshiyama, A. New one-dimensional conductors: graphitic microtubules. *Phys. Rev. Lett.* **68**, 1579–1581 (1992).
- Saito, R., Fujita, M., Dresselhaus, G. & Dresselhaus, M. S. Electronic structure of chiral graphene nanotubes. *Appl. Phys. Lett.* **60**, 2204–2206 (1992).
- Saito, R., Fujita, M., Dresselhaus, G. & Dresselhaus, M. S. Electronic structure of graphene nanotubes based on C_{60} . *Phys. Rev. B* **46**, 1804–1811 (1992).
- Zhang, Z. & Lieber, C. M. Nanotube structure and electronic properties probed by STM. *Appl. Phys. Lett.* **62**, 2972–2974 (1993).
- Olk, C. H. & Heremans, J. P. Scanning tunneling spectroscopy of carbon nanotubes. *J. Mater. Res.* **9**, 259–262 (1994).
- Ge, M. & Sattler, K. Vapor-condensation generation and STM analysis of fullerene tubes. *Science* **260**, 515–518 (1993).
- Carroll, D. L. *et al.* Electronic structure and localized states at carbon nanotube tips. *Phys. Rev. Lett.* **78**, 2811–2814 (1997).
- Ge, M. & Sattler, K. STM of single-shell nanotubes of carbon. *Appl. Phys. Lett.* **65**, 2284–2286 (1994).
- Thess, A. *et al.* Crystalline ropes of metallic carbon nanotubes. *Science* **273**, 483–487 (1996).
- Journet, C. *et al.* Large-scale production of single-walled carbon nanotubes by the electric-arc technique. *Nature* **388**, 756–758 (1997).
- Stroscio, J. A. & Feenstra, R. M. in *Scanning Tunneling Microscopy* (eds Stroscio, J. A. & Kaiser, W. J.) 95–141 (Academic, New York, 1993).
- Everson, M. P., Jaklevic, R. C. & Shen, W. Measurement of the local density of states on a metal surface: Scanning tunneling spectroscopic imaging of Au(111). *J. Vac. Sci. Technol. A* **8**, 3662–3665 (1990).
- Rao, A. M. *et al.* Diameter-selective Raman scattering from vibrational modes in carbon nanotubes. *Science* **275**, 187–191 (1997).
- Fischer, J. E. *et al.* Metallic resistivity in crystalline ropes of single-wall carbon nanotubes. *Phys. Rev. B* **55**, 4921–4924 (1997).
- Lee, R. S., Kim, H. J., Fischer, J. E., Thess, A. & Smalley, R. E. Conductivity enhancement in single-walled carbon nanotube bundles doped with K and Br. *Nature* **388**, 255–257 (1997).

Acknowledgements. We thank J. Liu and R. E. Smalley for discussions and samples, and D. Vezennov for help with the Au deposition. T.W.O. acknowledges fellowship support from the US NSF. This work was supported by the NSF Division of Materials Research.

Correspondence and requests for materials should be addressed to C.M.L. (e-mail: cml@cmliris.harvard.edu).

Deglacial changes in ocean circulation from an extended radiocarbon calibration

Konrad A. Hughen^{*†}, Jonathan T. Overpeck^{*‡}, Scott J. Lehman^{*}, Michael Kashgarian[§], John Southon[§], Larry C. Peterson^{||}, Richard Alley[¶] & Daniel M. Sigman[#]

^{*} INSTAAR and Department of Geological Sciences, University of Colorado, Boulder, Colorado 80309, USA

[‡] NOAA Paleoclimatology Program, NGDC, Boulder, Colorado 80303, USA

[§] CAMS, Lawrence Livermore National Laboratory, Livermore, California 94551, USA

^{||} RSMAS, University of Miami, Miami, Florida 33149, USA

[¶] Department of Geosciences & Earth System Science Center, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

[#] Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

[†] Present address: EPS Department, Harvard University, Cambridge, Massachusetts 02138, USA.

Temporal variations in the atmospheric concentration of radiocarbon sometimes result in radiocarbon-based age-estimates of biogenic material that do not agree with true calendar age. This problem is particularly severe beyond the limit of the high-resolution radiocarbon calibration based on tree-ring data, which stretches back only to^{1,2} about 11.8 kyr before present (BP), near the termination of the Younger Dryas cold period. If a wide range of palaeoclimate records are to be exploited for better understanding the rates and patterns of environmental change during the last deglaciation, extending the well-calibrated radiocarbon timescale back further in time is crucial. Several studies attempting such an extension, using uranium/thorium-dated corals^{3–5} and laminae counts in varved sediments^{6–9}, show conflicting results. Here we use radiocarbon data from varved sediments in the Cariaco basin, in the southern Caribbean Sea, to construct an accurate and continuous radiocarbon calibration for the period 9 to 14.5 kyr BP, nearly 3,000 years beyond the tree-ring-

based calibration. A simple model compared to the calculated atmospheric radiocarbon concentration and palaeoclimate data from the same sediment core suggests that North Atlantic Deep Water formation shut down during the Younger Dryas period, but was gradually replaced by an alternative mode of convection, possibly via the formation of North Atlantic Intermediate Water.

The annually laminated sediments of the Cariaco basin are composed of light (plankton-rich) and dark (mineral-rich) layers that result from strong seasonal fluctuations in trade-wind-induced upwelling and regional precipitation^{10,11}. Previous work has shown that relative reflectance (grey scale) and light-laminae thickness from the Cariaco basin are proxies for North Atlantic trade-wind intensity¹², which responds rapidly to changes in the temperature gradient between the high-latitude and tropical North Atlantic^{12–14}. It has been demonstrated previously that Cariaco basin light-laminae thickness and $\delta^{18}\text{O}$ from the GRIP Greenland ice core¹⁵ show similar event durations and patterns of change at the decade scale during the last deglaciation. This indicates that both responded to the same forcing (high-latitude North Atlantic sea surface temperature (SST)) and recorded synchronous events¹². In the present study, we investigated a sediment interval between 12.6 and 8.0 ¹⁴C kyr BP from core PL07-56PC containing thick, distinct laminations, and constructed a 5,500-year-long floating varve chronology by counting laminae pairs. From the ~8,000 ¹⁴C yr level to the sediment surface, laminae become thinner and less distinct. Monospecific samples (51) of the planktonic foraminifer *Globigerina bulloides* were taken for accelerator mass spectrometry (AMS) ¹⁴C dating, and an average measured reservoir age¹⁰ of 420 yr was subtracted from each date. Cariaco basin geometry dictates that only surface waters (<146 m) well-equilibrated with the atmosphere, and possessing a reservoir age of ~400 yr, can enter the basin, thus maintaining the basin reservoir age near open-ocean values. Despite a varying upwelling regime, ¹⁴C dating of multiple surface samples of known age indicates that the reservoir age of these samples agrees with present-day, open Atlantic values. Taken together, these lines of evidence suggest that the Cariaco basin reservoir age is not greatly influenced by local upwelling.

The 5,500-year-long floating varve chronology was anchored to an absolute calendar timescale by wiggle-matching¹⁶ radiocarbon versus calendar age variations to those generated from tree-ring

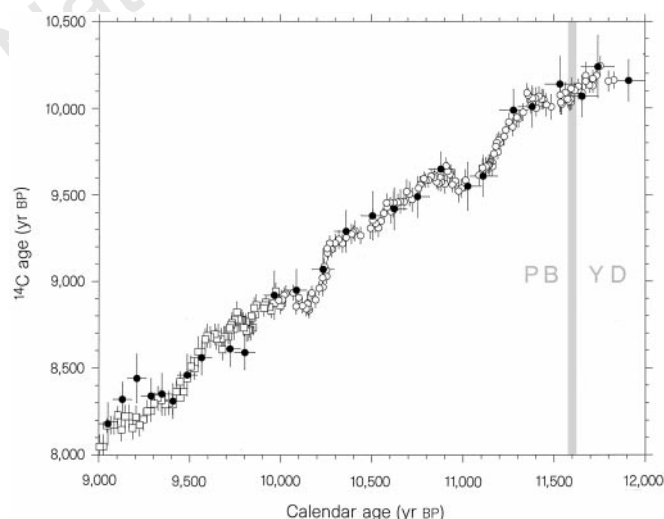


Figure 1 Variations in radiocarbon versus varve age for the Cariaco basin compared to those from tree-rings¹². Solid circles are data from Cariaco basin core PL07-56PC, open circles are from German pines and open squares are from German oaks. The relationship between German oak and pine-tree-ring data sets has been revised recently from that of Kromer and Becker¹ and Björck *et al.*², with the result that the pine chronology has now been securely anchored to the oak chronology (B. Kromer, personal communication). A 5,500-year floating chronology for the Cariaco basin was constructed by varve-counting between 12.6 and 8.0 ¹⁴C kyr BP. Approximately 3% of the varve-counted sediments contained short sections (1–2 cm thick) disturbed by microbioturbation that could not be bridged by switching to alternative cores. We interpolated through these sections using average varve thickness values from the immediately adjacent 10 cm of sediment above and below the disturbed sections. Maximum and minimum average varve thicknesses over 2-cm increments within the 10-cm sections were used to generate a cumulative uncertainty due to disturbance of 1.7% in 5,500 varve years. The match between the Cariaco basin and tree-ring data sets was carried out using a procedure that objectively minimizes the squared distance between independent time series¹⁶. The resulting correlation ($r = 0.993$) was used to anchor the floating Cariaco basin varve chronology to an absolute timescale at 9,051 yr BP. Inherent uncertainties in the Cariaco basin radiocarbon and varve ages introduce an estimated additional ± 50 yr uncertainty to the anchoring of the absolute varve chronology. Shaded line indicates the Younger Dryas (YD) to Preboreal (PB) transition as determined by Cariaco basin grey scale. ¹⁴C errors are shown at 2σ .

data¹ (Fig. 1). The correlation between the varve and tree-ring ¹⁴C variations is excellent ($r = 0.993$), and is used to anchor the varve chronology ending at 9051 calendar years before present (cal. yr BP). The data (Fig. 1) also show that the present-day reservoir age of 420 yr did not change during the interval 11.5–9.0 cal. kyr BP. In particular, the brief overlap of Cariaco basin and tree-ring ¹⁴C ages at the end of the Younger Dryas shows that the reservoir age remains unchanged even during times of greatly increased upwelling.

To assess the accuracy of the anchored varve chronology, we compared palaeoclimate records from the Cariaco basin and the GISP2 Greenland ice core. Cariaco basin grey scale and GISP2 accumulation^{17,18} are plotted against their independently derived annual timescales, from 15 to 9.0 cal. kyr BP (Fig. 2a). Based on an

objective method that minimizes the square of the distance between the two time-series¹⁶ (correlation $r = 0.667$), we transferred the GISP2 chronology to the Cariaco basin record and then directly compared the two annual chronologies (Fig. 2b). The two time-scales agree closely, well within the estimated errors, supporting the interpretation of similar palaeoclimate events in the two records as synchronous, as well as the accuracy of the Cariaco basin varve chronology.

The anchored Cariaco basin varve chronology was used to calibrate radiocarbon dates from 13 to 8.0 ¹⁴C kyr BP (15 to 9.0 cal. kyr BP; Supplementary Information; Fig. 3a, b). There are radiocarbon age plateaux centred at 11.7, 11.4 and 9.6 ¹⁴C kyr BP, in addition to a long, sloping 'plateau' from 10.6 to 10.0 ¹⁴C kyr BP,

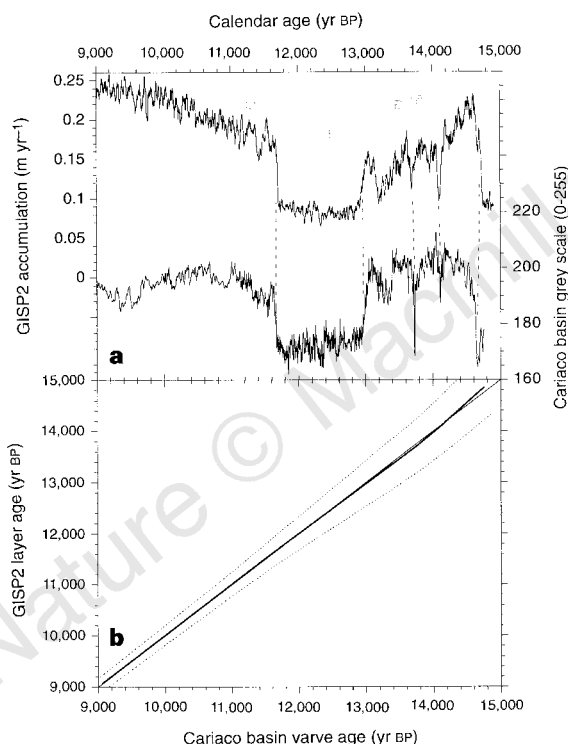


Figure 2 Independent assessment of Cariaco basin varve chronology using comparison to GISP2 palaeoclimate record. **a**, Cariaco basin grey scale¹² (lower curve) compared to accumulation data from the GISP2 ice core^{17,18} (upper curve) from 9.0 to 15 cal. kyr BP. Both data sets are plotted against their independent annual timescales. GISP2 accumulation is used here as a proxy for sub-polar North Atlantic SST based on the thermodynamic dependence of precipitation on temperature as well as its dynamic dependence on the location of storm tracks, which are controlled by North Atlantic SST gradient¹⁸. Large-scale, abrupt climate transitions and events that are identified as the same in Cariaco basin and GISP2 records, and used as tie points for correlation by a least-squares procedure¹⁶, are indicated by dashed lines. The correlation ($r = 0.669$) was used to assign GISP2 layer ages to the Cariaco basin grey scale record in order to compare the two chronologies (described in text). Letters indicate Preboreal (PB), Younger Dryas (YD), and Bølling/Allerød (B/A) climate periods. **b**, Cariaco basin varve age plotted versus GISP2 layer age (thick line). A one-to-one correlation is shown by the thin straight line and combined errors in the chronologies are indicated by the dotted lines. In places, the independent ages are nearly identical, and in all cases they agree well within estimated errors.

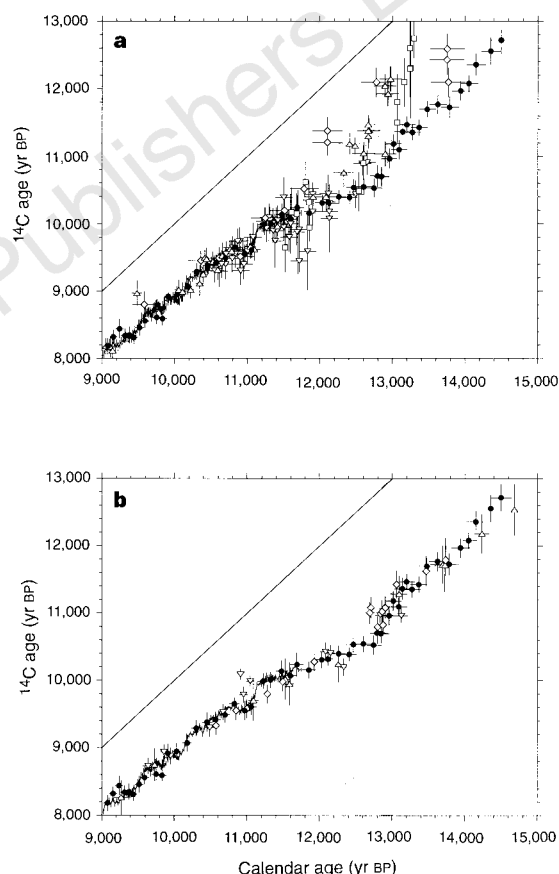


Figure 3 Radiocarbon ages versus calendar ages derived from new and published sources. **a**, Results from the Cariaco basin compared to those from European varved lake sediments and the Swedish varve chronology. Solid circles are Cariaco basin varves; solid line is German pine and German oak data¹; upside-down triangles (∇) are varves from Lake Gosiaz, Poland⁷; upright triangles (Δ) are varves from Lake Soppensee, Switzerland⁶; diamonds are varves from Lake Holzmaar, Germany⁸; open squares are Swedish varves⁹. **b**, Results from the Cariaco basin compared to those from coral U/Th dates from the Atlantic and Pacific oceans. Solid circles are Cariaco basin varves; solid line is German pine and German oak data¹; upside-down triangles are coral U/Th from Papua New Guinea⁴; upright triangles are coral U/Th from Barbados³; diamonds are coral U/Th from Tahiti⁵. Ten additional ¹⁴C dates from Cariaco basin core PL07-39PC (not shown for clarity) were correlated to PL07-56PC on the basis of distinct, millimetre-scale marker laminae. For each of the ten duplicate samples, the ages agree closely or are nearly identical, demonstrating the reproducibility of Cariaco basin ¹⁴C dates. Shaded lines indicate the beginning and end of the Younger Dryas based on Cariaco basin grey scale; line thickness indicates the length of these transitions in Cariaco basin varve years. All radiocarbon dates are presented with 2σ errors.

during which radiocarbon ages decrease only 600 yr in 1,600 varve years. The beginning of this sloping 'plateau' coincides with the onset of the climatic Younger Dryas, but continues for an additional 400 varve years after the Younger Dryas termination (Fig. 3a, b). Before 12 cal. kyr BP, our results generally agree with U/Th- ^{14}C data from Atlantic and Pacific corals³⁻⁵ (Fig. 3b), but disagree with calibration results from lakes Soppensee⁶ and Holzmaar⁸, and the Swedish varve chronology⁹ (Fig. 3a). As reported by Wohlfarth⁹, the Swedish chronology may contain some erroneous connections between varve segments. In addition, it appears that the Soppensee

and Holzmaar varve chronologies are missing varve years, possibly owing to the sediments being intermittently laminated during the Younger Dryas. This discrepancy cannot be attributed to a general land-sea difference due to a variable ocean reservoir age because the ocean radiocarbon dates, in this case, are younger than their implied calendar-age equivalents on land.

Calibrated radiocarbon ages were used to calculate radiocarbon activity ($\Delta^{14}\text{C}$) for the period 14.5–9.0 cal. kyr BP (Fig. 4). Although surface waters can be expected to smooth out year- to decade-long fluctuations in atmospheric $\Delta^{14}\text{C}$, comparisons of atmospheric and

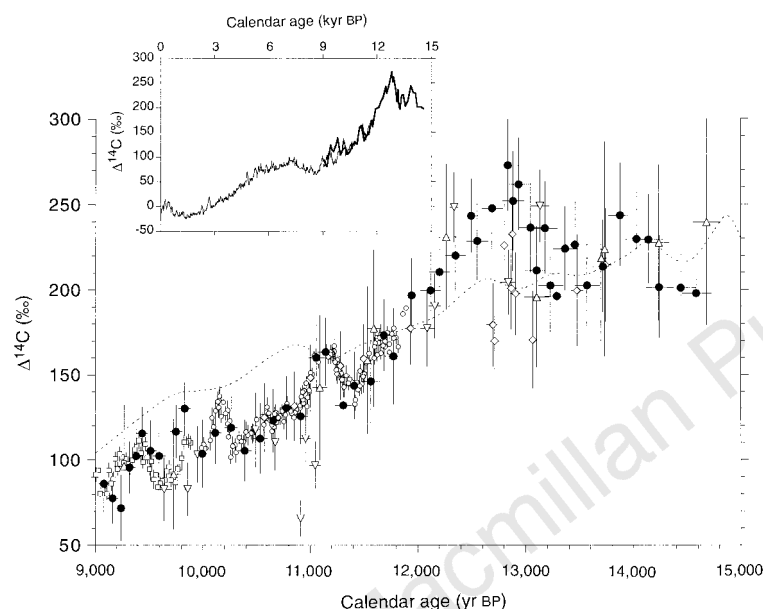


Figure 4 Atmospheric radiocarbon concentration ($\Delta^{14}\text{C}$), expressed as the difference in ^{14}C activity (measured in per mil) between the sample and a standard, after corrections for fractionation and sample age²⁹, calculated from various sources. Symbols are the same as in Fig. 3. $\Delta^{14}\text{C}_{\text{geo}}$, calculated from ^{14}C production rates resulting from changes in geomagnetic field intensity, is shown by dotted line. To ensure the intercomparability of results, the $\Delta^{14}\text{C}_{\text{geo}}$ values shown are from Goslar *et al.*⁷ using the calculations of Lal²⁹, and modelled using an initial $\Delta^{14}\text{C}$ of 230‰ at 15 kyr BP and field intensity data from Tric *et al.*³⁰. Inset shows $\Delta^{14}\text{C}$ data from the Cariaco basin with the continuous record from tree rings¹. Note increased amplitude of high-frequency fluctuations before ~10 cal. kyr BP.

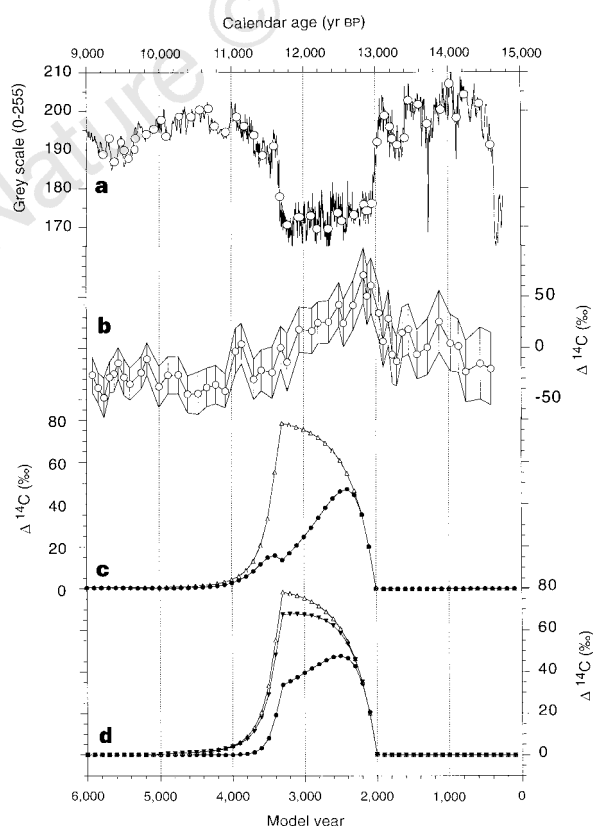


Figure 5 Observed palaeoclimate and $\Delta^{14}\text{C}$ from the Cariaco basin used to constrain box-model simulations of varying ocean circulation. **a**, Grey scale plotted versus calendar age for Cariaco basin core PL07-56PC. Open circles show locations of samples taken for AMS ^{14}C dating. **b**, Cariaco basin $\Delta^{14}\text{C}_{\text{res}}$ from core PL07-56PC corrected for geomagnetic production rate changes. Open circles are calculated values, upper and lower limits show 2σ errors. **c**, Simulated atmospheric $\Delta^{14}\text{C}$ from a geochemical box model as a function of different configurations of oceanic circulation. In the first experiment, NADW was switched abruptly from 20 sverdrups ($1 \text{ Sv} = 10^6 \text{ m}^3 \text{ s}^{-1}$) to 0 Sv at model year 2000, and then abruptly from 0 to 20 Sv at model year 3330 (open triangles). In a second experiment (solid circles), NADW was specified as above, with the addition of NAIW formation and export increasing linearly from 0 to 20 Sv during the period from model year 2200 to 3330, then abruptly switching back to 0 Sv at 3330 (timing of peak ^{14}C activity depends on arbitrary lag in NAIW initiation). Model circulation regimes were prescribed as changes from the control run of ref. 24. NAIW is formed in the boreal Atlantic surface box and exported according to the following scheme: 25% is recirculated within the North Atlantic; 75% is incorporated into an 'intermediate conveyor' and advected into Pacific and Indian Ocean intermediate boxes. **d**, Same as **c**, with the first experiment simulating a shut-down of NADW (open triangles). In the second experiment, instead of NAIW formation and export, convection in the Southern Ocean was increased gradually by 0 to 20 Sv during the Younger Dryas (solid triangles). For the third experiment (solid circles), in addition to increased convection, gas exchange rates for the surface Southern Ocean were increased by 2x throughout the run, and the mass exchange rates between the deep Southern Ocean and surrounding deep ocean boxes were increased during the Younger Dryas by 40 Sv each (4x total exchange).

oceanic time series of bomb ^{14}C activity suggest that surface waters faithfully record past changes in atmospheric $\Delta^{14}\text{C}$ lasting just a few decades¹⁹. In Fig. 4 calculated $\Delta^{14}\text{C}$ is compared to results from tree rings^{1,2} and corals^{3–5}, as well as estimated $\Delta^{14}\text{C}$ arising from changes in ^{14}C production due to shifts in Earth's geomagnetic field strength ($\Delta^{14}\text{C}_{\text{geo}}$), as previously modelled by Goslar *et al.*⁷. To isolate the changes in atmospheric $\Delta^{14}\text{C}$ arising from exchange between carbon reservoirs (the additional possibility of forcing from solar variability is rejected because proxy data show no large fluctuations in solar activity at that time; B. Finkel, personal communication), we subtracted $\Delta^{14}\text{C}_{\text{geo}}$ from observed $\Delta^{14}\text{C}$ to obtain a residual ($\Delta^{14}\text{C}_{\text{res}}$). Comparing this to a grey-scale record from the same Cariaco basin core (Fig. 5a, b) allows a direct correlation of North Atlantic palaeoclimate and atmospheric $\Delta^{14}\text{C}$ proxy records. Our data reveal a large asymmetrical increase in $\Delta^{14}\text{C}_{\text{res}}$ during the Younger Dryas, with a rise of 50–80‰ occurring in ~200 years, followed by a gradual decline of the same amplitude occurring over 1,100 years. The rapid rise in $\Delta^{14}\text{C}_{\text{res}}$ is synchronous, within errors, with the onset of the climatic Younger Dryas, suggesting that changes in both $\Delta^{14}\text{C}_{\text{res}}$ and climate were caused by the same forcing mechanism.

It has been suggested previously that shifts in North Atlantic thermohaline circulation were a possible cause of rapid changes in $\Delta^{14}\text{C}$ in addition to climate during the Younger Dryas^{20–23}. To test this, we used a geochemical box model of the oceans and atmosphere²⁴ to simulate atmospheric $\Delta^{14}\text{C}$ as a function of changes in oceanic circulation. In the first experiment, North Atlantic Deep Water (NADW) formation was switched abruptly 'off', kept 'off' for the duration of the Younger Dryas, and was returned abruptly 'on' at the Younger Dryas termination. As seen in some earlier studies^{2,7}, the simulated $\Delta^{14}\text{C}$ increase (80‰) is relatively gradual, followed by an abrupt decrease at the Younger Dryas termination (Fig. 5c, d), whereas the observed time series of $\Delta^{14}\text{C}_{\text{res}}$ peaks early and declines gradually throughout the Younger Dryas (Fig. 5b). A simple shut-down of NADW seems to explain the increase but not the subsequent drawdown of $\Delta^{14}\text{C}_{\text{res}}$ during the Younger Dryas, implying the presence of an additional mechanism which lowers atmospheric $\Delta^{14}\text{C}$ during the interval of reduced deep-water formation.

Several published general circulation model (GCM) experiments involving freshwater perturbations to the North Atlantic that interrupted or stopped deep-water production also show a spontaneous oceanic response in the form of either increased North Atlantic Intermediate Water (NAIW) formation or increased convection in the Southern Ocean^{25,26}. Box-model experiments reveal that NAIW formation and export to other ocean basins (in the manner of NADW today) provides an effective sink for atmospheric ^{14}C . Using the same schedule of NADW formation as in the first experiment, we can reproduce observed $\Delta^{14}\text{C}_{\text{res}}$ by gradually increasing NAIW formation and export during the period in which NADW has been shut down (Fig. 5b, c). Similar experiments show that simply increasing convection in the Southern Ocean during the period in which NADW is shut down has very little effect on modelled $\Delta^{14}\text{C}$ (Fig. 5d), primarily due to incomplete isotopic equilibration of surface waters, and inadequate export of deep Southern Ocean water to other deep basins. Doubling the gas exchange rate and quadrupling the overall exchange with adjoining deep-water reservoirs results in a ^{14}C drawdown during the Younger Dryas (Fig. 5d), but it is smaller than that seen either in observations (Fig. 5b) or in model experiments with NAIW formation and export (Fig. 5c).

Ocean GCMs indicate that NAIW formation is much less effective at bringing heat north than NADW²⁶, although we find it can be an effective sink for atmospheric ^{14}C . It is thus possible that a shutdown of NADW produced dramatic cooling of the North Atlantic region and increased atmospheric $\Delta^{14}\text{C}$ during the Younger Dryas, while its gradual replacement by NAIW acted to draw down $\Delta^{14}\text{C}$ even as cooling was maintained. This model is compatible with observa-

tional constraints ($\Delta^{14}\text{C}$ and palaeoclimate) from the Cariaco basin, and is also in agreement with previous studies suggesting some form of continued North Atlantic sub-surface ventilation during the Younger Dryas²⁷, which may have occurred specifically through the formation of NAIW^{21,28}. In addition to a detailed ^{14}C calibration for the entire interval between 14.5 and 11.5 cal. kyr BP, our data provide a clear picture of how atmospheric $\Delta^{14}\text{C}$ changed during the last deglaciation, and place needed constraints on the origin of atmospheric ^{14}C changes and possible modes of ocean circulation during the Younger Dryas. □

Received 7 November 1996; accepted 17 November 1997.

- Kromer, B. & Becker, B. German oak and pine ^{14}C calibration, 7200–9400 BC. *Radiocarbon* **35**, 125–136 (1993).
- Björck, S. *et al.* Synchronized terrestrial-atmosphere deglacial records around the North Atlantic. *Science* **274**, 1155–1160 (1996).
- Bard, E., Arnold, M., Fairbanks, R. G. & Hamelin, B. ^{230}Th – ^{234}U and ^{14}C ages obtained by mass spectrometry on corals. *Radiocarbon* **35**, 191–199 (1993).
- Edwards, R. L. *et al.* A large drop in atmospheric $^{14}\text{C}/^{12}\text{C}$ and reduced melting in the Younger Dryas, documented with ^{230}Th ages of corals. *Science* **260**, 962–967 (1993).
- Bard, E. *et al.* Deglacial sea-level record from Tahiti corals and the timing of global meltwater discharge. *Nature* **382**, 241–244 (1996).
- Hajdas, I. *et al.* AMS radiocarbon dating and varve chronology of Lake Sopensee: 6000 to 1200 ^{14}C years BP. *Clim. Dyn.* **9**, 107–116 (1993).
- Goslar, T. *et al.* High concentration of atmospheric ^{14}C during the Younger Dryas cold episode. *Nature* **377**, 414–417 (1995).
- Hajdas, I. *et al.* AMS radiocarbon dating of annually laminated sediments from Lake Holzmaar, Germany. *Quat. Sci. Rev.* **14**, 137–143 (1995).
- Wohlfarth, B. The chronology of the last termination: a review of radiocarbon-dated, high-resolution terrestrial stratigraphies. *Quat. Sci. Rev.* **15**, 267–284 (1996).
- Hughen, K. A., Overpeck, J. T., Peterson, L. C. & Anderson, R. F. In *Palaeoclimatology and Palaeoceanography from Laminated Sediments* (ed. Kemp, A. E. S.) 171–183 (Spec. Publ. 116, Geol. Soc., London, 1996).
- Peterson, L. C., Overpeck, J. T., Kipp, N. G. & Imbrie, J. A high-resolution late Quaternary upwelling record from the anoxic Cariaco Basin, Venezuela. *Paleoceanography* **6**, 99–119 (1991).
- Hughen, K. A., Overpeck, J. T., Peterson, L. C. & Trumbore, S. Rapid climate changes in the tropical Atlantic region during the last deglaciation. *Nature* **380**, 51–54 (1996).
- Overpeck, J. T., Peterson, L. C., Kipp, N., Imbrie, J. & Rind, R. Climate change in the circum-North Atlantic region during the last deglaciation. *Nature* **338**, 553–557 (1989).
- Rind, D., Peteet, D., Broecker, W. S., McIntyre, A. & Ruddiman, W. The impact of cold North Atlantic sea surface temperatures on climate: implications for the Younger Dryas cooling (11–10 k). *Clim. Dyn.* **1**, 33 (1986).
- Johnsen, S. J. *et al.* Irregular glacial interstadials recorded in a new Greenland ice core. *Nature* **359**, 311–313 (1992).
- Paillard, D. Macintosh program makes time-series analysis easy. *Eos* **77**, 379 (1996).
- Alley, R. B. *et al.* Abrupt increase in Greenland snow accumulation at the end of the Younger Dryas event. *Nature* **362**, 527–529 (1993).
- Kapsner, W. R., Alley, R. B., Schuman, C. A., Anandakrishnan, S. & Grootes, P. M. Dominant influence of atmospheric circulation on snow accumulation in Greenland over the past 18,000 years. *Nature* **373**, 52–54 (1995).
- Druffel, E. M. Decade time scale variability of ventilation in the North Atlantic: high-precision measurements of bomb radiocarbon in banded corals. *J. Geophys. Res.* **94**, 3271–3285 (1989).
- Broecker, W. S., Peteet, D. M. & Rind, D. Does the ocean-atmosphere system have more than one stable mode of operation? *Nature* **315**, 21–26 (1985).
- Boyle, E. A. & Keigwin, L. D. North Atlantic thermohaline circulation during the past 20,000 years linked to high-latitude surface temperature. *Nature* **330**, 35–40 (1987).
- Stocker, T. F. & Wright, D. G. Rapid changes in ocean circulation and atmospheric radiocarbon. *Paleoceanography* **11**, 773–795 (1996).
- Mikolajewicz, U. A meltwater induced collapse of the thermohaline circulation and its influence on the oceanic distribution of $\delta^{14}\text{C}$ and $\delta^{18}\text{O}$. 1–25 (Tech. Rep. 189, Max-Planck-Inst. für Meteorologie, Hamburg, 1996).
- Keir, R. S. On the late Pleistocene ocean geochemistry and circulation. *Paleoceanography* **3**, 413–445 (1988).
- Manabe, S. & Stouffer, R. J. Simulation of abrupt climate change induced by freshwater input to the North Atlantic Ocean. *Nature* **378**, 165–167 (1995).
- Rahmstorf, S. Rapid climate transitions in a coupled ocean-atmosphere model. *Nature* **372**, 82–85 (1994).
- Charles, C. D. & Fairbanks, R. G. Evidence from Southern Ocean sediments for the effect of North Atlantic deep-water flux on climate. *Nature* **355**, 416–419 (1992).
- Lehman, S. J. & Keigwin, L. D. Sudden changes in North Atlantic circulation during the last deglaciation. *Nature* **356**, 757–762 (1992).
- Stuiver, M. & Pollach, H. A. On the reporting of ^{14}C ages. *Radiocarbon* **19**, 355–359 (1977).
- Tric, E. *et al.* Paleointensity of the geomagnetic field during the last 80,000 years. *J. Geophys. Res.* **97**, 9337–9351 (1992).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) and as paper copy from Mary Sheehan at the London editorial office of *Nature* and also at the World Data Center-A for Paleoclimatology (<http://www.ngdc.noaa.gov>).

Acknowledgements. We thank S. Trumbore and M. Kerwin for comments that improved the manuscript; B. Kromer and S. Björck for sharing new data for German pine and oak tree-ring chronologies; T. Goslar and E. Bard for sharing numerical data from $\Delta^{14}\text{C}$ modelling results; H.-L. Lin for ^{14}C dates from core PL07-39PC; and J. Moore for help making thin sections. This work was supported by the US NSF and the National Oceanic and Atmospheric Administration, as well as the US Department of Energy through Lawrence Livermore National Laboratory, and by NSF funding to the Lamont-Doherty Earth Observatory Deep-Sea Sample Repository.

Correspondence and requests for materials should be addressed to K.A.H. (e-mail: hughen@fas.harvard.edu).

The effect of loading rate on static friction and the rate of fault healing during the earthquake cycle

Chris Marone

Department of Earth, Atmospheric, and Planetary Sciences, MIT, Cambridge, Massachusetts 02139, USA

The seismic cycle requires that faults strengthen (heal) between earthquakes, and the rate of this healing process plays a key role in determining earthquake stress drop^{1–4}, rupture characteristics^{5,6} and seismic scaling relations^{2–4,7}. Frictional healing (as evidenced by increasing static friction during quasi-stationary contact between two surfaces^{1,8–12}) is considered the mechanism most likely to be responsible for fault strengthening^{2,3,13,14}. Previous studies, however, have shown a large discrepancy between laboratory and seismic (field) estimates of the healing rate^{2–4,14,15}; in the laboratory, rock friction changes by only a few per cent per order-of-magnitude change in slip rate, whereas seismic stress drop increases by a factor of 2 to 5 per order-of-magnitude increase in earthquake recurrence interval. But in such comparisons, it is assumed that healing and static friction are independent of loading rate. Here, I summarize laboratory measurements showing that static friction and healing vary with loading rate and time, as expected from friction theory^{16–18}. Applying these results to seismic faulting and accounting for differences in laboratory, seismic and tectonic slip rates, I demonstrate that post-seismic healing is expected to be retarded for a period of several hundred days following an earthquake, in agreement with recent findings from repeating earthquakes^{13,14,19,20}.

The problem of how faults regain strength between earthquakes is crucial for our understanding of the seismic cycle and the physics of earthquake rupture. Several factors could play a role, such as interaction of geometric irregularities and fault-zone heterogeneity⁷, but frictional strengthening is likely to be a key effect in any case. Seismic data indicate that earthquake stress drop increases logarithmically with time^{2,3,14} and rock friction studies show that static friction (which would correspond to fault yield strength) increases with log(time) (refs 1, 9–12). Critical factors in understanding these observations are the connection between stress drop and fault strength, and the problem of scaling laboratory friction observations to earthquake faults. Previous studies have assumed that static friction data can be applied directly, whereas the results summarized here indicate that static friction and time-dependent strengthening are system responses, dependent on loading rate and elastic coupling.

Figure 1 shows friction data from experiments in which simulated fault gouge (breccia and wear material within fault zones) was sheared in a servocontrolled testing apparatus. Frictional healing $\Delta\mu$ is measured as the difference between static and initial friction; the data indicate that static friction and $\Delta\mu$ increase with hold time and loading rate (Fig. 1), consistent with previous work^{1,8,9,18}. The experiments include instantaneous changes in loading rate (Fig. 1), which show that the effect of loading rate on static friction is significantly larger than on steady-state sliding friction. Temporal characteristics of the data and measurements of changes in relative porosity of the gouge $\Delta\phi$ indicate that shear strength decays logarithmically with time during healing and that gouge compacts with the same temporal form (Fig. 2), consistent with previous work^{21,22}. The data indicate that static friction and healing rate

($\beta = \Delta\mu/\Delta\log_{10}t_h$, where t_h is hold time) vary systematically with loading rate over a range of t_h spanning three orders of magnitude (Fig. 1). In particular, a ten times increase in loading rate has about the same effect on $\Delta\mu$ as a ten times increase in hold time.

The data have three important implications. (1) They indicate that laboratory measurements of static friction and time-dependent healing must be scaled appropriately for comparison with seismic estimates of fault healing. Previous comparisons have assumed that fault healing can be derived directly from the steady-state velocity dependence of sliding friction^{2–4,15}. (2) The observation that restrengthening is a system response, dependent on loading velocity, indicates that the rate of fault healing is expected to vary during the seismic cycle. (3) The observation that porosity decreases with

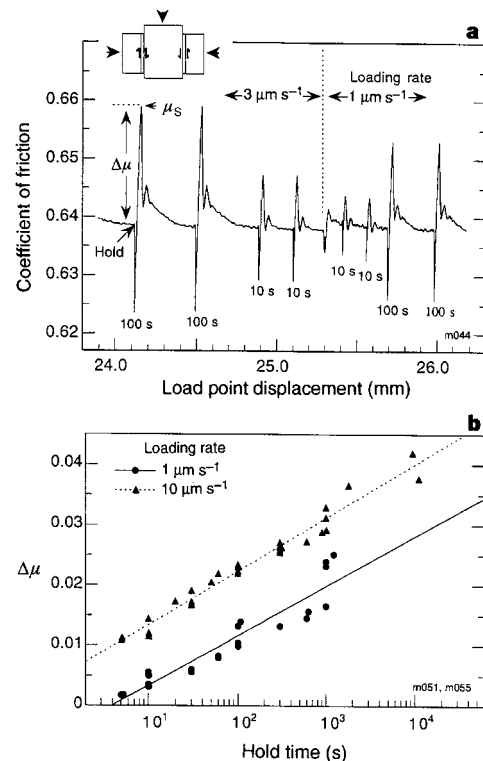


Figure 1 Friction data from experiments in which gouge layers were sheared within rough granite surfaces in the double-direct-shear configuration (upper inset). **a**, Plot of the coefficient of friction versus displacement of a control point where shear load is applied. Load point displacement is controlled with a high-speed, servo-hydraulic actuator. Data are shown for slide-hold-slide tests at loading rates of 1 and 3 $\mu\text{m s}^{-1}$ and hold times of 10 and 100 s. The coefficient of static friction, μ_s , is the maximum value reached following a hold, and $\Delta\mu$ is the difference between static and sliding friction. Hold times given below data indicate positions at which loading was stopped and fault surfaces were held in quasi-stationary contact. Frictional strength relaxes during holds due to fault creep and finite apparatus stiffness. Inset shows double-direct-shear configuration. The testing apparatus consists of two independently servocontrolled load frames applying shear and normal forces across the sample. Stiffness of the vertical frame is 225 MPa cm^{-1} expressed as shear stress per load point displacement. Nominal frictional contact dimension is 10 cm $\times$ 10 cm and gouge layers are initially 2.1 mm thick. Initial gouge particle size is 50–150 μm and rock surface roughness (r.m.s.) $\sim$ 200 μm . During shearing, normal stress is held constant at 25 MPa and changes in layer thickness are recorded continuously to $\pm 0.1 \mu\text{m}$. **b**, Frictional healing data for two loading rates as a function of hold time. The data indicate that $\Delta\mu$ and β are functions of hold time and loading rate, with higher loading rates yielding greater healing for a given time. Lines indicate best-fit log-linear relations. Healing rates $\beta = \Delta\mu/\Delta\log_{10}t_h$ are: 0.0082 for 1 $\mu\text{m s}^{-1}$, and 0.0086 for 10 $\mu\text{m s}^{-1}$.

log(hold time), coupled with the dependence of both $\Delta\phi$ and $\Delta\mu$ on loading rate, implies that porosity is closely related to the evolution of frictional strength as suggested in theoretical and experimental studies^{21–23}.

An important point is the origin of the loading-rate dependence of $\Delta\mu$ and β : why does static frictional strength vary with loading rate? The observation that healing varies approximately as the product of loading rate and hold time implies that the mechanism is a function of time and slip. As this is a property of rate- and state-dependent friction (RSF) laws, which are capable of modelling a wide range of seismic phenomena^{4,18,24–27}, I evaluate the present data in the context of these laws. In RSF relations^{16,17} the coefficient of friction μ is written as:

$$\mu = \mu_0 + a \ln\left(\frac{V}{V_0}\right) + b \ln\left(\frac{V_0\theta}{D_c}\right) \quad (1)$$

$$\frac{d\theta}{dt} = 1 - \frac{V\theta}{D_c} \quad (2)$$

where μ_0 is a constant appropriate for steady-state slip at velocity V_0 , V is the frictional slip rate, D_c is a critical slip distance, θ is a state variable, and a and b are empirical constants. Equation (2) describes the evolution of frictional state following a perturbation in velocity as proposed by Dieterich¹⁶. Other evolution laws have been proposed^{16,17} but the distinction is not important here. To model laboratory friction data, equations (1) and (2) are coupled with a description of elastic interaction between the loading system and sliding surfaces:

$$d\mu/dt = k(V_1 - V) \quad (3)$$

where k is stiffness divided by normal stress, and V_1 is loading velocity. Equations (1)–(3) are solved numerically, taking $V_1 = V_0$, to yield friction curves for comparison with data.

In Fig. 3a I show a best-fit RSF model to a step change in loading rate. The main features of the data are well fitted, and the best-fit constitutive parameters are consistent with previous studies^{9,28–30}. These parameters were used to obtain predictions of static friction and healing as a function of hold time and velocity (Fig. 3b). The predicted curves match the data for each loading rate, which is remarkable given that both $\Delta\mu$ and β are fitted for a set of loading rates with no free parameters. RSF laws are apparently capable of describing frictional strengthening as observed in experiment, and thus may be used to scale the laboratory results to different loading velocities.

In experiments the initial slip velocity and the subsequent loading rate are equal and varied together, whereas on earthquake faults the initial slip velocity would correspond, crudely, to the coseismic slip rate V_{cs} and the rate of subsequent loading, V_L , to the background tectonic loading rate. To estimate the effect of slip rate on fault healing I performed numerical simulations using equations (1)–(3) (Fig. 4a). The simulations use laboratory values for the parameters a and b , a field-based estimate for D_c , and $k = G/(L\sigma)$, where G is shear modulus, L is earthquake rupture dimension, and σ is the fault-normal effective stress. A reference case, in which $V_{cs} = V_L = 1 \mu\text{m s}^{-1}$, indicates log-linear healing, as expected (Fig. 4a). To compare other cases, I consider only the change in static friction by subtracting the steady-state velocity effect on friction, given by $b - a = \Delta\mu/\ln V$. This is reasonable if friction during seismic rupture is limited by dynamic factors and is independent of static failure strength; however, only the absolute values of $\Delta\mu$ are affected in any case, not its time dependence. For coseismic slip rates of $0.01\text{--}1 \text{ m s}^{-1}$ and loading rates approximately equal to plate tectonic rates ($30 \text{ mm yr}^{-1} \approx 3 \times 10^{-9} \text{ m s}^{-1}$) static friction shows an initial period of retarded healing and an increase in healing rate after a few hundred days (Fig. 4a). Although the

length of the initial period of delayed healing varies somewhat with stiffness (and thus earthquake size), the form of the curve does not.

In the simulations I used $k = 1 \times 10^{-5} \mu\text{m}^{-1}$ (which corresponds to $L = 30 \text{ m}$ for $G = 30 \text{ GPa}$ and $\sigma = 100 \text{ MPa}$) for comparison with data from a set of small repeating earthquakes (Fig. 3b). The earthquakes are magnitude 1.5 events that repeatedly ruptured a small patch (radius $\approx 30 \text{ m}$, depth $6\text{--}8 \text{ km}$) of the Calaveras fault in California, with a range of inter-event times spanning nearly three orders of magnitude^{13,14}. The data indicate that seismic stress drop increases with recurrence interval, and that the healing rate increases at about 100–300 days, in agreement with the numerical simulation (Fig. 4). Moreover, a reduced healing rate during the period immediately following earthquakes of similar size can be inferred from repeating earthquakes with shorter recurrence interval ($<10\text{--}100$ days), which do not show variation in stress drop with recurrence interval^{19,20}.

In the numerical simulations, the change in healing rate is the result of slip during frictional relaxation. This slip is akin to earthquake afterslip, in which post-seismic fault slip accumulates at a rate that decays approximately exponentially following some

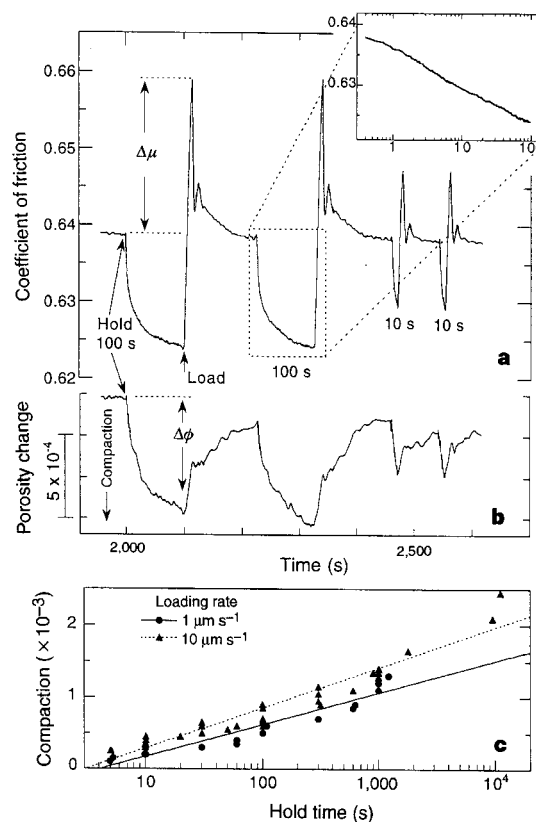
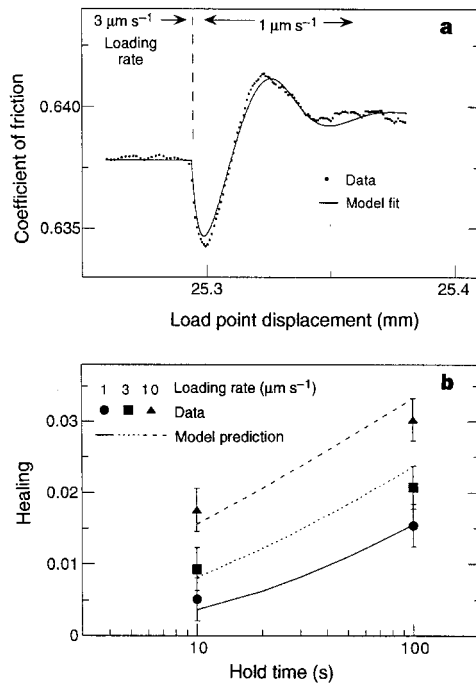


Figure 2 Experimental determination of variation of $\Delta\phi$ and $\Delta\mu$. **a** and **b**, plots of friction data (**a**) and relative porosity (**b**) versus time for the initial section of Fig. 1. Frictional strength decreases, and the gouge layers compact, during holds. The magnitude of $\Delta\phi$ and $\Delta\mu$ increase with hold time. Dilation of the gouge layers on reloading approximately equals time-dependent compaction, and thus porosity during steady-state shear is constant. Upper inset shows friction versus log(time) during hold and indicates that friction relaxes logarithmically with time. Porosity mimics friction variations during holds. Porosity data are obtained from measurements of layer thickness, using the nominal contact area to determine volume change, which is normalized by the initial gouge volume. A linear trend has been removed from the porosity data to account for geometric thinning of the gouge layer during shear. **c**, Compaction $\Delta\phi$ associated with healing is shown versus hold time for two loading rates. Like the frictional healing data, $\Delta\phi$ is a function of hold time and loading rate, with higher loading rates yielding greater healing for a given time. Lines indicate best-fit log-linear relations.

seismic events^{24,31}. The RSF laws show (equation (2)) that afterslip would tend to erase the effects of frictional healing until slip velocity $V \ll D_c/\theta$. Thus for hold times that are short compared with the time necessary for afterslip to slow sufficiently—about 100 days in the simulations—frictional healing is partially erased (Fig. 4a). This same effect produces concave-up curvature in laboratory data for short hold times (shorter than those shown here; Fig. 1), which is consistent with the view that strengthening involves chemical cementation and contact junction growth, and is disrupted by slip.



The agreement between the simulations and the repeating-earthquake data implies that similar processes may operate in earthquake fault zones. But when considering the rough agreement between the long-term healing rates of Fig. 4 it must be noted that changes in static strength are a function of both β and effective normal stress. I have used laboratory β values reported here and a normal effective lithostatic gradient, which yields higher effective normal stress than some field estimates⁴. Lower normal stress would require higher frictional healing rates, which is consistent with enhancement of

Figure 3 Use of an RSF model to analyse loading-rate dependence. **a**, Friction data for a step decrease in loading rate (as shown in Fig. 1a) together with a best-fit RSF model determined by iterative least-squares inversion of equations (1)–(3). The best-fit constitutive parameters (see text) are: $a = 0.0066$, $b = 0.0083$, $D_c = 7.1 \mu\text{m}$. These data show velocity weakening frictional behaviour (the steady-state coefficient of sliding friction decreases with slip velocity). Complete stress–displacement curves for similar experiments²⁹ show that sliding friction evolves to a steady value over 3–6 mm of slip and undergoes a transition from velocity strengthening to velocity weakening at engineering shear strains of 4–5. The critical slip distance, D_c (a measure of the slip or shear strain needed to reach a new steady-state friction upon a perturbation in slip rate) also evolves over this interval²⁹. **b**, Measurements of healing ($\Delta\mu$) for three loading rates, together with prediction of the rate and state friction law. Model predictions use the constitutive parameters determined in **a** and thus the dependence of $\Delta\mu$ and β on loading rate and hold time are fitted with no free parameters. Error bars show typical experimental reproducibility.

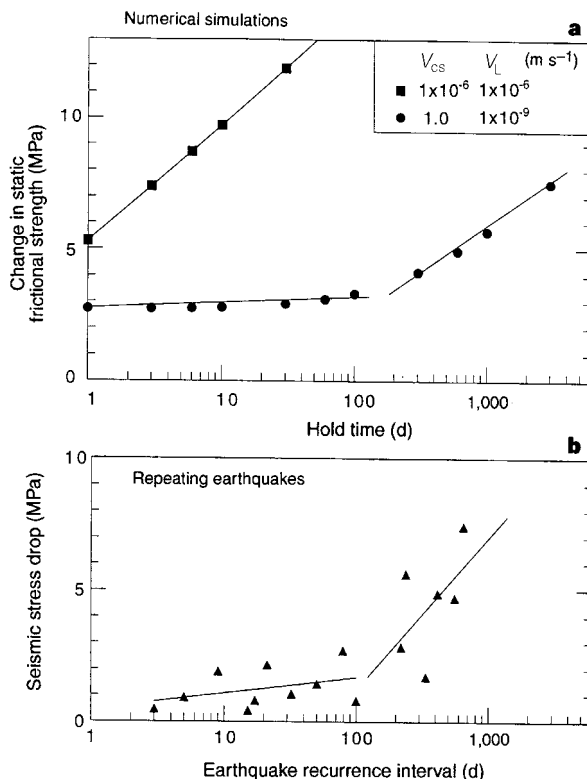


Figure 4 Results of numerical simulations and comparison with field data. **a**, Numerical simulations of fault healing in which initial slip velocity (corresponding to coseismic slip) V_{cs} and subsequent loading rate (corresponding to tectonic loading) V_L are varied. In this model, hold time corresponds to earthquake recurrence interval. Frictional strength is derived from static friction assuming an effective fault normal stress of 100 MPa. For $V_{cs} = V_L = 1 \mu\text{m s}^{-1}$, static strength increases linearly with $\log(\text{time})$. Taking V_{cs} as 1 m s^{-1} and V_L as 10^{-9} m s^{-1} (typical values for earthquake particle velocity and plate tectonic loading rates, respectively) indicates that the effective healing rate is retarded for the period immediately following an earthquake. The duration of reduced healing scales with fault stiffness and thus earthquake rupture dimension, which is taken as 30 m. Friction parameters are $a = 0.01$, $b = 0.02$, and $D_c = 5 \text{ mm}$, consistent with laboratory and field estimates²⁹. **b**, Seismic stress drop versus recurrence interval for a series of repeating earthquakes^{13,14}. The data indicate a distinct break in slope, reflecting a lower healing rate for events with recurrence interval $< \approx 100$ days. In the numerical simulations shown in **a** such variations in healing rate are due to afterslip. In both panels lines indicate best-fit log–linear relations for times < 100 days and > 100 days.

healing by thermal and chemical effects^{4,11,12}. But even for effective fault-normal stress as low as 10 MPa, the simulations of Fig. 4 do not indicate the apparent discrepancy between laboratory and seismic estimates of fault healing noted in earlier works^{2-4,15}. This is because those works used an effective slip velocity, derived from D_c and the earthquake recurrence interval, which is not a correct approach³¹. Frictional healing rate β scales as $\ln t_h$ for long hold times (from equations (1) and (2), $V \ll D_c/\theta$ as t_h becomes large, which yields $d\theta/dt \approx 1$ and thus $\Delta\mu(t_h)$ proportional to $\ln t_h$), whereas the approach based on effective velocity assumes that healing and fault strength scale as $\Delta\mu(t_h) \approx (b-a)\ln t_h$. This indicates that the comparison of laboratory and field estimates of fault healing should involve absolute values rather than relative changes. Using this approach, I note that the average healing rate from repeating earthquakes¹⁴ and seismic scaling relations²⁻⁴ is ~ 3 MPa per decade and the long-term rate from repeating earthquakes (Fig. 4b) is ~ 6 MPa per decade. This compares with the long-term healing rate inferred from laboratory data of 4.5 MPa per decade (Fig. 4a), depending on b and effective normal stress, which is in much better agreement than previous comparisons. □

Received 29 May; accepted 21 October 1997.

- Dieterich, J. H. Time-dependent friction in rocks. *J. Geophys. Res.* **77**, 3690–3697 (1972).
- Scholz, C. H., Aviles, C. A. & Wesnousky, S. G. Scaling differences between large interplate and intraplate earthquakes. *Bull. Seismol. Soc. Am.* **76**, 65–70 (1986).
- Kanamori, H. & Allen, C. R. in *Earthquake Source Mechanics* (eds Das, S., Boatwright, J. & Scholz, C.) 227–236 (Am. Geophys. Un., Washington DC, 1986).
- Scholz, C. H. *The Mechanics of Earthquakes and Faulting* (Cambridge Univ. Press, New York, 1990).
- Heaton, T. Evidence for and implications of self healing pulses of slip in earthquake rupture. *Phys. Earth Planet. Inter.* **64**, 1–20 (1990).
- Perrin, G., Rice, J. R. & Zheng, G. Self-healing slip pulse on a frictional surface. *J. Mech. Phys. Solids* **43**, 1461–1495 (1995).
- Heimpel, M. Critical behaviour and the evolution of fault strength during earthquake cycles. *Nature* **388**, 856–858 (1997).
- Johnson, T. Time dependent friction of granite: implications for precursory slip on faults. *J. Geophys. Res.* **86**, 6017–6028 (1981).
- Beeler, N. M., Tullis, T. E. & Weeks, J. D. The roles of time and displacement in the evolution effect in rock friction. *Geophys. Res. Lett.* **21**, 1987–1990 (1994).
- Nakatani, M. & Mochizuki, H. Effects of shear stress applied to surface in stationary contact on rock friction. *Geophys. Res. Lett.* **23**, 869–872 (1996).
- Karner, S. L., Marone, C. & Evans, B. Laboratory study of fault healing and lithification in simulated fault gouge under hydrothermal conditions. *Tectonophysics* **277**, 41–55 (1997).
- Olsen, M. P., Scholz, C. H. & Léger, A. Healing and sealing of a simulated fault gouge under hydrothermal conditions: implications for fault healing. *J. Geophys. Res.* (submitted).
- Vidale, J. E., Ellsworth, W., Cole, A. & Marone, C. Rupture variation with recurrence interval in eighteen cycles of a small earthquake. *Nature* **368**, 624–626 (1994).
- Marone, C., Vidale, J. E. & Ellsworth, W. Fault healing inferred from time dependent variations in source properties of repeating earthquakes. *Geophys. Res. Lett.* **22**, 3095–3098 (1995).
- Cao, T. & Aki, K. Effect of slip rate on stress drop. *Pure Appl. Geophys.* **124**, 515–529 (1986).
- Dieterich, J. H. Modeling of rock friction: 1. Experimental results and constitutive equations. *J. Geophys. Res.* **84**, 2161–2168 (1979).
- Ruina, A. Slip instability and state variable friction laws. *J. Geophys. Res.* **88**, 10359–10370 (1983).
- Kato, N., Yamamoto, K., Yamamoto, H. & Hirasawa, T. Strain-rate effects on frictional strength and the slip nucleation process. *Tectonophysics* **211**, 269–282 (1992).
- Nadeau, R. M., Foxall, W. & McEvilly, T. V. Clustering and periodic recurrence of microearthquakes on the San Andreas fault at Parkfield, California. *Science* **267**, 503–508 (1995).
- Lees, J. M. Microseismic multiplets do not exhibit effects of fault healing at Coso. *Eos* **77**, 502 (1996).
- Wang, W. & Scholz, C. H. Micromechanics of the velocity and normal stress dependence of rock friction. *Pure Appl. Geophys.* **143**, 303–316 (1994).
- Beeler, N. M. & Tullis, T. E. The roles of time and displacement in velocity dependent volumetric strain of fault zones. *J. Geophys. Res.* **102**, 22595–22609 (1997).
- Segall, P. & Rice, J. R. Dilatancy, compaction, and slip instability of a fluid infiltrated fault. *J. Geophys. Res.* **100**, 22155–22173 (1995).
- Marone, C., Scholz, C. H. & Bilham, R. On the mechanics of earthquake afterslip. *J. Geophys. Res.* **96**, 8441–8452 (1991).
- Rice, J. R. Spatio-temporal complexity of slip on a fault. *J. Geophys. Res.* **98**, 9885–9907 (1993).
- Dieterich, J. H. A constitutive law for rate of earthquake production and its application to earthquake clustering. *J. Geophys. Res.* **99**, 2601–2618 (1994).
- Roy, M. & Marone, C. Earthquake nucleation on models faults with rate and state dependent friction: the effects of inertia. *J. Geophys. Res.* **101**, 13919–13932 (1996).
- Biegel, R. L., Sammis, C. G. & Dieterich, J. H. The frictional properties of a simulated gouge with a fractal particle distribution. *J. Struct. Geol.* **11**, 827–846 (1989).
- Marone, C. & Kilgore, B. Scaling of the critical slip distance for seismic faulting with shear strain in fault zones. *Nature* **362**, 618–621 (1993).
- Beeler, N. M., Tullis, T. E., Blanpied, M. L. & Weeks, J. D. Frictional behaviour of large displacement experimental faults. *J. Geophys. Res.* **101**, 8697–8716 (1996).
- Marone, C. Laboratory-derived friction laws and their application to seismic faulting. *Annu. Rev. Earth Planet. Sci.* (in the press).

Acknowledgements. I thank C. Scholz and N. Beeler for review comments and J. Rice, M. Blanpied and C. Raymo for discussions that helped sharpen my understanding of frictional healing. This work was supported by the US NSF.

Correspondence should be addressed to the author (e-mail: cjm@westerly.mit.edu).

Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest

J. H. Lawton*, D. E. Bignell†‡, B. Bolton‡, G. F. Bloemers*‡‡, P. Eggleton‡, P. M. Hammond‡, M. Hodda‡‡, R. D. Holt§, T. B. Larsen||, N. A. Mawdsley*‡‡, N. E. Stork‡‡, D. S. Srivastava*‡‡ & A. D. Watt¶

* NERC Centre for Population Biology, Imperial College, Silwood Park, Ascot SL5 7PY, UK

† School of Biological Sciences, Queen Mary and Westfield College, University of London, London E1 4NS, UK

‡ The Natural History Museum, Cromwell Road, London SW7 5BD, UK

§ Museum of Natural History, University of Kansas, Lawrence, Kansas 66045, USA

|| 358 Coldharbour Lane, London, SW9 8LP, UK

¶ Institute of Terrestrial Ecology, Bush Estate, Penicuik, Midlothian, EH26 0QB, UK

Despite concern about the effects of tropical forest disturbance and clearance on biodiversity^{1,2}, data on impacts, particularly on invertebrates, remain scarce³⁻⁸. Here we report a taxonomically diverse inventory on the impacts of tropical forest modification at one locality. We examined a gradient from near-primary, through old-growth secondary and plantation forests to complete clearance, for eight animal groups (birds, butterflies, flying beetles, canopy beetles, canopy ants, leaf-litter ants, termites and soil nematodes) in the Mbalmayo Forest Reserve, south-central Cameroon. Although species richness generally declined with increasing disturbance, no one group serves as a good indicator taxon⁹⁻¹² for changes in the species richness of other groups. Species replacement from site to site (turnover) along the gradient also differs between taxonomic groups. The proportion of 'morphospecies' that cannot be assigned to named species and the number of 'scientist-hours' required to process samples both increase dramatically for smaller-bodied taxa. Data from these eight groups indicate the huge scale of the biological effort required to provide inventories of tropical diversity, and to measure the impacts of tropical forest modification and clearance.

We chose birds and butterflies as groups to study because they frequently serve as 'flagship taxa' in biodiversity inventories^{6,9,13}. The other six groups were chosen because they are all widespread and abundant invertebrates. The study area¹⁴⁻¹⁶ is semi-deciduous humid forest representative of much of southern Cameroon, and is used for subsistence farming, experimental agriculture, and forestry trials¹⁶, providing a mosaic of habitats. Study sites (Table 1) formed a gradient of increasing intensity and frequency of disturbance. For some taxa the inventories are almost complete; for others, because of lower sampling effort, they are far from complete (Table 2). The proportion of 'morphospecies'¹¹ (judged by expert systematists) that cannot currently be assigned to known (named) species is inversely related to log (body length) ($F_{1,5} = 11.07$, $P = 0.02$, ants included only once; values $> x$ in Table 2 regressed as x), emphasizing the huge disparity that exists between groups in the state of taxonomic knowledge.

Generally, as in previous studies^{1,2,5-7,17,18}, species richness

Present addresses: TBCU, University of Malaysia Sabah, 88999 Kota Kinabalu Sabah, Malaysia (D.E.B.); Central Science Laboratory, Plant Health Group, Sand Hutton, York YO4 1LZ, UK (G.F.B.); Australian National Insect Collection, CSIRO Division of Entomology, GPO Box 1700, Canberra, Australia (M.H.); Department of Biology, University of Leeds, Leeds, LS2 9JT, UK (N.A.M.); Cooperative Research Centre for Tropical Rainforest Ecology and Management, PO Box 6811, Cairns, Queensland 4870, Australia (N.E.S.); and Centre for Biodiversity Research, Faculty of Science, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada (D.S.S.)

declined with increasing habitat modification (Fig. 1), although richness in some groups was surprisingly insensitive to extreme habitat modification, and not all taxa had maximum species richness in near-primary (NP) or old-growth secondary (OS) sites. No single group is a good indicator of changes in species richness for all, or even most, other groups (Table 3). Although sample sizes for the correlations in Table 3 are small, the number of positive, statistically significant correlations (5 in 45) is very low (to emphasize the point, we have not corrected for multiple comparisons). On average, only about 10–11% of the variation (mean $r = 0.33$) in species richness of one group is predicted by the change in richness of another group. Most r -values are positive, and 19 of the 45 exceed 0.5, but 10 correlations are actually negative (one significantly so)^{11,12}. Data for birds and butterflies are uncorrelated. Changes in richness of only one group (canopy ants) are significantly positively correlated with changes in richness of three others; most groups are significantly positively correlated with only one or two others. Beetles caught in interception traps are not correlated with any other group. Malaise-trap beetles tend to be negatively correlated with other groups, significantly so with nematodes. One of the five significant positive correlations in Table 3 is between the same taxon (ants), sampled in the canopy and in the litter, but canopy beetles and understorey flying beetles are not correlated. Overall, the distribution of correlation coefficients in Table 3 appears idiosyncratic.

We conclude that attempts to assess the impacts of tropical forest modification and clearance using changes in the species richness of one or a limited number of indicator taxa^{9–12} (including popular groups such as birds or butterflies) to predict changes in richness of other taxa may be highly misleading. Several other studies also cast serious doubt on the utility of indicator species to predict changes in species richness^{11,12,19,20}, presumably for the common-sense reason that different kinds of organisms (birds and nematodes, to take an extreme example) have very different ecological requirements, and hence are extremely unlikely to show similar responses to even major changes in habitat.

Even if richness changes little with disturbance, trophic structure

Table 1 Study sites in the Mbalmayo Forest Reserve

Near primary (NP)	Residual plots of near-primary forest, apparently lightly and selectively logged >70 years ago. Sample plots 2–3 ha in extent, surrounded by old-growth secondary forest.
Old-growth secondary (OS)	The dominant vegetation over large parts of south-central Cameroon; large tracts of second-growth forest, unlogged for >40 years. Sample sites (~2 ha) were representative of many parts of the Forest Reserve. Local tree species richness in one 22-ha sample plot was 149 spp.; data are not available for other parts of the forest.
Partial manual clearance, plus plantation (PManC)	1-ha plots. Felling and/or poisoning of up to 70% of old second-growth canopy trees, and understorey clearance in rows with machete. <i>Terminalia ivorensis</i> planted in 1987–88, at 5–8 m intervals, and 10–15 m tall at time of sampling. <i>T. ivorensis</i> is native to Cameroon, but does not grow wild in Mbalmayo. Resulting forest has closed canopy of native trees and <i>T. ivorensis</i> , with a dense understorey.
Partial mechanical clearance, plus plantation (PMechC)	1-ha plots. Bulldozer used to clear ~50% of understorey and larger old second-growth trees, resulting in some soil compaction. <i>T. ivorensis</i> planted in 1987–88, as above. More disturbed, and with less understorey than PManC.
Complete clearance, with young plantation (CC)	1-ha plots. All large trees felled with a chainsaw, smaller trees, remaining vegetation, and dead wood cleared manually and mechanically by bulldozer. Considerable soil compaction. Row-planted with <i>T. ivorensis</i> (1–2 m tall at time of sampling); considerable bare ground and no closed canopy. For some taxa (such as termites and nematodes) this is a more extreme disturbance than the farm fallow (which lacks small trees) because topsoil is removed.
Manually cleared farm fallow (FF)	A ~3 ha plot, cleared manually of trees and other vegetation in 1990; some wood from felled trees spread and allowed to decompose. Weeded to prevent tree regeneration; ground cover of dense <i>Chromolaena</i> . Similar fallow fields created by local farmers occur throughout the forest.

may alter, and species characteristic of primary and old-growth secondary forest may be replaced by species associated with disturbed habitats^{2,6,7,17,18,21}. With some exceptions^{14,15}, too little is known about the ecology of most of the taxa to comment in detail, although birds characteristic of primary forest disappear, and species from disturbed habitats are added along the gradient (Fig. 1a). A modified version of Whittaker's β (β -2; see Methods) quantifies species' replacement²² (Table 2); higher values imply greater changes in species composition with disturbance. Rates of turnover appear to be very different in different groups.

The richness and composition of the regional species pool²³ may account for some of these patterns. For example, 66–80% of the bird species on the most severely disturbed plots (complete clearance (CC) and manually cleared farm fallow (FF)) do not occur in the forest, but have entered the area from natural grasslands and savannas well to the north. Because the Mbalmayo region currently lacks savanna butterflies and termites (which might otherwise colonize the severely disturbed plots), there is less species replacement, and more impact of severe disturbance on species richness, than might be observed at sites with forest and savanna species in close proximity.

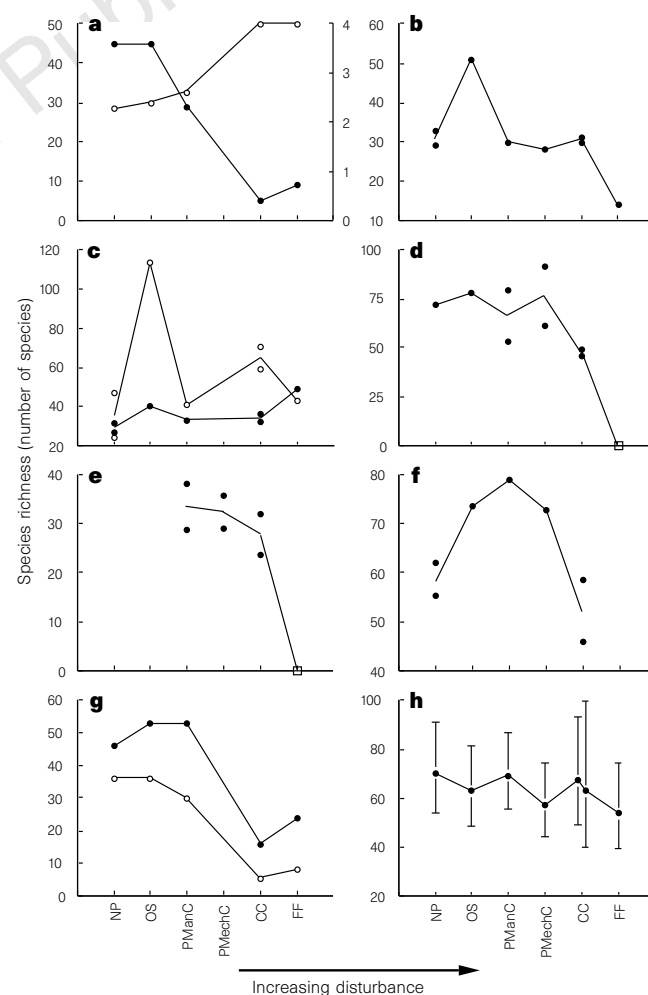


Figure 1 Species richness of animal groups along a gradient of increasing habitat modification (see Table 1). **a**, Birds (with mean habitat scores (open circles) on right ordinate); **b**, butterflies; **c**, flying beetles (malaise traps, filled circles; flight-interception traps, open circles); **d**, canopy beetles; **e**, canopy ants (open squares in **d** and **e** assume richness falls to zero in the absence of canopy); **f**, leaf-litter ants; **g**, termites (open circles, additional surveys not included in analyses); **h**, soil nematodes (with 95% confidence, and data from two even more heavily disturbed sites¹⁴ not included in analyses).

Table 2 Important features of the taxa and inventory

Group surveyed	Total species recorded on study plots during survey*	Index of species turnover (β -2)	Geometric mean body length (range) (cm)	Scientist-hours to sample, sort and identify	Morphospecies that cannot be assigned to known species (%)
Birds (Aves)	78 (H)	18.3	26.5 (9–78)	50	0
Butterflies (Lepidoptera)	132 (I)	22.7	4.7 (2.0–11.1)	150	1
Flying beetles (Coleoptera)	358 (malaise) (L)	45.0	0.37 (0.06–2.26)	600 (both methods combined)	50–70
Canopy beetles (Coleoptera)	467 (interception) 342 (L)	22.1 39.4	0.45 (0.05–4.0)	1,000	>80
Canopy ants (Hymenoptera: Formicidae)	96 (I)	30.4	0.32 (0.1–1.0)	160	40
Leaf-litter ants (Hymenoptera: Formicidae)	111 (I)	6.8	0.32 (0.1–1.0)	160	40
Termites (Isoptera)	114 (H)	28.8	0.38 (0.2–0.75)	2,000	30
Soil nematodes (Nematoda)	374 (I)	21.0	0.089 (0.02–0.4)	6,000	>90

* H (high, >90%), I (intermediate, 50–90%) and L (low, <50%) are estimates of the degree of completeness in the species inventory for each group.

Table 3 Correlations between groups in change in species richness across plots on the disturbance gradient

	Butterflies	Flying beetles (malaise)	Flying beetles (interception)	Canopy beetles	Canopy ants	Leaf-litter ants	Termites	Nematodes
Birds	5, 0.67 <i>P</i> = 0.21	5, <i>-0.40</i> <i>P</i> = 0.50	5, 0.28 <i>P</i> = 0.65	4, 0.98 <i>P</i> = 0.02 (5, 0.78 <i>P</i> = 0.12)	2, – (3, 0.50 <i>P</i> = 0.66)	4, 0.47 <i>P</i> = 0.53	5, 0.90 <i>P</i> = 0.04	5, 0.51 <i>P</i> = 0.38
Butterflies		5, <i>-0.31</i> <i>P</i> = 0.61	5, 0.82 <i>P</i> = 0.09	5, 0.40 <i>P</i> = 0.50 (6, 0.75 <i>P</i> = 0.09)	3, <i>-0.49</i> <i>P</i> = 0.67 (4, 0.97 <i>P</i> = 0.03)	5, 0.25 <i>P</i> = 0.69	5, 0.58 <i>P</i> = 0.30	6, 0.41 <i>P</i> = 0.42
Flying beetles (malaise traps)			5, 0.24 <i>P</i> = 0.71	4, 0.25 <i>P</i> = 0.75 (5, <i>-0.75</i> <i>P</i> = 0.15)	2, – (3, <i>-0.99</i> <i>P</i> = 0.08)	4, 0.43 <i>P</i> = 0.57	5, <i>-0.32</i> <i>P</i> = 0.61	5, <i>-0.98</i> <i>P</i> = 0.006
Flying beetles (interception traps)				4, 0.28 <i>P</i> = 0.72 (5, 0.37 <i>P</i> = 0.54)	2, – (3, 0.30 <i>P</i> = 0.81)	4, 0.21 <i>P</i> = 0.80	5, 0.21 <i>P</i> = 0.74	5, <i>-0.17</i> <i>P</i> = 0.79
Canopy beetles					3, 0.86 <i>P</i> = 0.33 (4, 0.97 <i>P</i> = 0.03)	5, 0.67 <i>P</i> = 0.22	4, 0.92 <i>P</i> = 0.08 (5, 0.73 <i>P</i> = 0.17)	5, <i>-0.28</i> <i>P</i> = 0.66 (6, 0.57 <i>P</i> = 0.24)
Canopy ants						3, 0.99 <i>P</i> = 0.01	2, – (3, 0.46 <i>P</i> = 0.70)	3, 0.04 <i>P</i> = 0.98 (4, 0.70 <i>P</i> = 0.30)
Leaf-litter ants							4, 0.84 <i>P</i> = 0.16	5, <i>-0.21</i> <i>P</i> = 0.73
Termites								5, 0.50 <i>P</i> = 0.38

Where two estimates of species richness for a group exist at a particular point on the disturbance gradient (for example, butterflies in NP; Fig. 1b), values have been averaged. Data are then presented as sample size (number of plots available for pairwise comparison, from 2 to 6), Pearson's r , and associated probability, P . Values in parentheses are calculated using an assumed species richness of zero for canopy-inhabiting taxa in the absence of a canopy. Ten negative r -values and associated probabilities are in italics; P -values <0.05 are in bold. To simplify presentation, r -values are rounded to two significant figures, and hence similar values in the table (but not in the original calculations) may be associated with different P -values. Dashes indicate sample sizes too small for correlations to be calculated.

The number of scientist-hours required to process these samples (Table 2) is inversely related to each group's geometric mean body length ($\log_{10}(h) = -0.72 \log_{10}(cm) + 2.64$; $F_{1,5} = 11.64$, $P = 0.02$, including ants only once). In total, approximately 5 scientist-years (over 10,000 scientist-hours; Table 2) were required to sample, sort and catalogue the approximately 2,000 species. The total number of species present in all taxa must be at least 10 times, and may approach 100 times, this number, because many of the inventories are incomplete (Table 2) and numerous taxa (many very species rich, and many very small bodied) were not included in the surveys. We conclude that a detailed inventory, even from one area, of the impacts of tropical forest modification and disturbance on biodiversity will require a huge scientific effort, far exceeding anything attempted so far anywhere in the world. These data are also informative about the scale of effort required to produce an all-taxa biological inventory (ATBI)^{11,24} for a 'representative hectare' of forest²⁵. Such an inventory is unlikely to be possible without 1–2 orders of magnitude more effort than that devoted to the present

surveys. There were 12 systematists working on our material; two orders of magnitude more effort might absorb 10–20% of the estimated entire global workforce of 7,000 systematists²⁶ if an ATBI were to be completed in a reasonable time.

We do not argue against the critical importance of undisturbed forest for many species. Maximum diversity in tropical forests will be conserved by maintaining a mosaic of habitats, including large tracts of primary and old-growth secondary forest^{27,28}. Our data do, however, suggest that some managed forests (plantations of native tree species, established by partial forest clearance^{17,27}) may help to maintain diversity for both invertebrate and vertebrate²⁸ taxa. Most importantly, assessing the effects of habitat modification and disturbance on tropical biodiversity by using changes in species richness of familiar and well-studied groups (such as birds or butterflies) as indicators of changes in the diversity of other taxa gives a highly misleading picture of overall faunal changes. Monitoring changes in biodiversity resulting from forest modification and destruction requires a wide range of taxa to be studied,

embracing species with very different ecologies and life histories. The resources required to do this far exceed those currently available to taxonomists and ecologists worldwide. □

Methods

Study area. Mbalmayo Forest Reserve (11° 25' to 11° 31' E, 3° 23' to 3° 31' N) is ~50 km south of Yaoundé in Cameroon, and at 650 m altitude. We did not establish experimental treatments ourselves (Table 1). We used existing areas of forest, silviculture demonstration plots created by the Institute of Terrestrial Ecology (UK) and later the Forest Management and Regeneration Project (a bilateral Government of Cameroon ONADEP, and UK Overseas Development Administration project), and experiments established by the International Institute of Tropical Agriculture field station¹⁶. Sites selected were representative of different parts of the forest habitat mosaic; replicate samples (Fig. 1) were taken from different sites. Human impacts have extirpated some species (including several large mammals) completely from the reserve; however, most other well-known groups (birds, butterflies and termites) remain rich in characteristic forest species.

Sampling. Sampling, concentrated in 1992–94, minimized edge effects by locating surveys and traps away from habitat boundaries. Within a group, sampling effort was identical on all plots (standardized by time, number of samples, or number of individuals sampled). Although some of the plots are relatively small (and hence data may not be representative of similar treatments applied over much larger areas), and although sampling of some groups is much less complete than others, our primary focus is on patterns of change along the gradient, not on absolute changes in species richness; the data are adequate for these purposes. Authors primarily responsible for data for each group are indicated in the following by initials.

Birds. (R.D.H., J.H.L.) Plots were censused visually and aurally, in morning or late-afternoon sessions lasting 1–2 h (10 h per plot, 7 h in July 1993 and 3 h in March 1994). Only birds using the plots (not flying over or through) were counted. Species 'characteristic habitat scores'²⁹ were assigned as follows: 1, 'true forest', primary forest, old-growth secondary forest; 2, 'forest'; 3, forest and other habitats, such as bush, second growth, clearings; 4, bush, scrub, farmland, gardens.

Butterflies. (T.B.L., D.S.S., N.E.S., A.D.W.) Plots were sampled in March 1994 (ref. 16) by netting by D.S.S., netting by four hired local collectors, rotated round plots to avoid collector bias, and by fruit-baited traps.

Flying beetles. (P.E., P.M.H., N.A.M.) Two methods, malaise and flight-interception traps³⁰, caught beetles from just above ground level to a maximum height of ~1 m (two traps in NP and CC plots, one in each OS, PmanC and FF plot; see Table 2 for definitions). Each trap was set for two periods of 4 days (July 1993). The two methods catch very different components of the flying-beetle fauna.

Canopy beetles and canopy ants. (B.B., N.S.E., A.D.W.) Sampling¹⁶ was confined to planted *Terminalia ivorensis* in plantations (15 trees per plot) and to indigenous *T. superba* in NP and OS plots, by fogging with permethrin, an insecticide with rapid knock-down (ants, on four occasions between 1991 and November 1993; beetles in November 1993 only). There was no surviving tree canopy on the FF plot; canopy-insect species richness was assigned values of zero on this plot for some analyses.

Leaf-litter ants. (B.B., N.E.S., A.D.W.) Ten litter samples, each 1 m², taken in 50-m transects across each plot in November 1993. Leaf litter and the top few millimetres of soil were sieved in a coarse 1-cm sieve, and the residue extracted in 'Winkler bags'¹⁶.

Termites. (D.E.B., P.E., J.H.L.) All species (up to 1 m above ground) were collected from transects (100 m × 2 m) run across each plot in November 1992 and July 1994 (ref. 16). Additional less comprehensive samples¹⁵ (August 1992 and July 1993) are shown (Fig. 1) for comparative purposes only.

Nematodes. (G.F.B., M.H., J.H.L.) Core samples¹⁴ taken from 18 sites in July 1993 comprise 200 nematodes identified to species in one core from each site, with similar sites grouped to obtain replicate measures (three each of NP, OS, PMechC and CC; four of PManC; and two of FF).

Body size. We used geometric mean body length (for butterflies, wingspan), hereafter referred to as 'length', of the largest and smallest species in each group (adult birds, beetles and nematodes, worker ants and termites) as an estimate of each group's characteristic body size.

Sampling effort. 'Scientist-hours' (Table 2) include fieldwork, extraction, sorting to morphospecies, identification where possible, and data compilation. They exclude training, travel, and domestic and logistic activities. Estimates (made for each taxon without consulting colleagues working on other groups) are minima, because sorting and identification involved world experts with direct access to NHM collections. Conversion from scientist-hours to scientist-years assumes 250 working days of 8 h per year.

Statistics. Data are total species recorded per plot, without errors; replicated plots, where available (Fig. 1), give similar results. For nematodes, 95% confidence intervals were calculated using replicate plots¹⁴. Where the total number of individuals sampled is known we also used a jack-knife technique (the program EstiMates; R. K. Colwell, University of Connecticut, unpublished) to calculate mean richness and 95% confidence intervals for data rarefied to uniform numbers of individuals sampled on all plots, and re-ran the analyses. Our general conclusions are unaffected. Pearson correlations (Table 3) were calculated with SPSS 6.1. β -2 (species' turnover along the gradient)²² = $\{[(S/\alpha_{\max}) - 1]/(N - 1)] \times 100$, where S is the total number of species recorded, $\alpha_{\max}$ is the maximum within-taxon richness per sample and N is the number of samples per taxon. β -2 is insensitive to trends in species richness, and allows comparison between transects of unequal size²². It ranges from 0 (no turnover) to 100 (every sample has a unique set of species). Because values reflect real turnover and pseudo-turnover (caused by incomplete inventories), and cannot be compared statistically, they must be interpreted cautiously; different ways²² of calculating β , and of including or excluding duplicate samples, do not alter the general conclusions.

Received 30 May; accepted 2 October 1997.

- Whitmore, T. C. & Sayer, J. A. (eds) *Tropical Deforestation and Species Extinction* (Chapman and Hall, London, 1992).
- Marshall, A. G. & Swaine, M. D. Tropical rain forest: disturbance and recovery. *Phil. Trans. R. Soc. Lond. B* **335**, 323–457 (1992).
- Sutton, S. L. & Collins, N. M. in *Conservation of Insects and their Habitats* (eds Collins, N. M. & Thomas, J. A.) 405–424 (Academic, London, 1991).
- Collins, N. M. The effect of logging on termite diversity and decomposition processes in lowland dipterocarp forests. *Trop. Ecol. Dev.* **198**, 113–121 (1980).
- Wood, T. G. et al. Abundance and distribution of termites (Isoptera) in a riparian forest in the southern Guinea savanna vegetation zone of Nigeria. *Biotropica* **14**, 25–39 (1982).
- Bowman, D. J. M. S. et al. Slash-and-burn agriculture in the wet coastal lowlands of Papua New Guinea: responses of birds, butterflies and reptiles. *J. Biogeog.* **17**, 227–239 (1990).
- Hill, J. K. et al. Effects of selective logging on tropical forest butterflies on Buru, Indonesia. *J. Appl. Ecol.* **32**, 754–760 (1995).
- Watt, A. D. et al. in *Forests and Insects* (eds Watt, A. D., Stork, N. E. & Hunter, M. D.) 271–284 (Chapman and Hall, London, 1997).
- Davidar, P. et al. in *Measuring and Monitoring Biodiversity in Tropical and Temperate Forests* (eds Boyle, T. J. B. & Boontawe, B.) 287–301 (Centre for International Forestry Research, Bogor, Indonesia, 1995).
- Gaston, K. J. (ed.) *Biodiversity: A Biology of Numbers and Difference* (Blackwell, Oxford, 1996).
- Oliver, I. O. & Beattie, A. J. Designing a cost-effective invertebrate survey: a test of methods for rapid assessment of biodiversity. *Ecol. Appl.* **6**, 594–607 (1996).
- Prendergast, J. R. & Eversham, B. C. Species richness covariance in higher taxa: empirical tests of the biodiversity indicator concept. *Ecography* **20**, 210–216 (1997).
- Louette, M. & Bijnens, L. The utility of birds as bioindicators—case-studies in equatorial Africa. *Belg. J. Zool.* **125**, 157–165 (1995).
- Bloemers, G. F. et al. The effects of forest disturbance on diversity of tropical soil nematodes. *Oecologia* **111**, 575–582 (1997).
- Eggleton, P. et al. The diversity, abundance and biomass of termites under differing levels of disturbance in the Mbalmayo Forest Reserve, southern Cameroon. *Phil. Trans. R. Soc. Lond. B* **351**, 51–68 (1996).
- Watt, A. D. et al. Impact of forest management on insect abundance and damage in a lowland tropical forest in southern Cameroon. *J. Appl. Ecol.* **34**, 985–998 (1997).
- Estrada, A., Coates-Estrada, R. & Merritt, D. Jr Non flying mammals and landscape changes in the tropical rain forest region of Lost Tustlas, Mexico. *Ecography* **17**, 229–241 (1994).
- Canaday, C. Loss of insectivorous birds along a gradient of human impact in Amazonia. *Biol. Conserv.* **77**, 63–77 (1996).
- Prendergast, J. R. et al. Rare species, the coincidence of diversity hotspots and conservation strategies. *Nature* **365**, 335–337 (1993).
- Dobson, A. P. et al. Geographic distribution of endangered species in the United States. *Science* **275**, 550–553 (1997).
- Johns, A. G. Bird population persistence in Sabahan logging concessions. *Biol. Conserv.* **75**, 3–10 (1996).
- Harrison, S., Ross, S. J. & Lawton, J. H. Beta diversity on geographic gradients in Britain. *J. Anim. Ecol.* **61**, 151–158 (1992).
- Cornell, H. V. & Lawton, J. H. Species interactions, local and regional processes, and limits to the richness of ecological communities: a theoretical perspective. *J. Anim. Ecol.* **61**, 1–12 (1992).
- Eggleton, P. & Bignell, D. E. in *Insects in a Changing Environment* (eds Harrington, R. & Stork, N. E.) 473–497 (Academic, London, 1995).
- May, R. M. How many species are there on earth? *Science* **241**, 1441–1449 (1988).
- Blackmore, S. Knowing the Earth's biodiversity: challenges for the infrastructure of systematic biology. *Science* **274**, 63–64 (1996).
- Blockhus, J. M. et al. (eds) *Conserving Biological Diversity in Managed Tropical Forests. Proc. Workshop, IUCN General Assembly, Perth, Australia 1990* (Gland, IUCN/ITTO, Cambridge, 1992).
- Johns, A. G. & Johns, B. G. Tropical forests and primates: long term co-existence. *Oryx* **29**, 205–211 (1995).

29. Mackworth-Præd, C. W. & Grant, C. H. B. *Birds of West Central and Western Africa. African Handbook of Birds Series 3*, Vols I and II (Longman, London, 1981).
30. Hammond, P. M. in *Insects and the Rain Forests of South East Asia (Wallacea)* (eds Knight, W. J. & Holloway, J. D.) 197–254 (Royal Entomological Society, London, 1990).

Acknowledgements. We thank T. Blackburn, R. Booth, M. Brendell, T. Brown, J. Davies, R. Didham, D. Nguete, J. Fanguem, G. Lawson, R. Leakey, E. Lindsay, J. Mason, M. Mboglen, C. McBeath, P. Ngehi, A. Roby, R. Sands, Z. Tchoundjeu, J. Tchoupa, J. Tipa, F. Wanless, B. Waite, M. West, T. Wood and E. Wright for technical and taxonomic help and discussion. J.H.L. is an honorary research associate at the NHM. We thank the Government of Cameroon (Office National de Développement des Forêts), the International Institute of Tropical Agriculture, the UK Overseas Development Administration, and the UK Natural Environment Research Council TIGER Programme for logistical and financial support.

Correspondence and requests for materials should be addressed to J.H.L. (e-mail: j.lawton@ic.ac.uk).

A neuronal network for computing population vectors in the leech

John E. Lewis* & William B. Kristan Jr

Department of Biology, University of California, San Diego, La Jolla, California 92093-0357, USA

The correlation of neuronal activity with sensory input and behavioural output has revealed that information is often encoded in the activity of many neurons across a population, that is, a neural population code is used^{1,2}. The possible algorithms that downstream networks use to read out this population code have been studied by manipulating the activity of a few neurons in a population^{3,4}. We have used this approach to study population coding in a small network underlying the leech local bend, a body bend directed away from a touch stimulus⁵. Because of the small size of this network we are able to monitor and manipulate the complete set of sensory inputs to the network. We show here that the population vector⁶ formed by the spike counts of the active mechanosensory neurons is well correlated with bend direction. A model based on the known connectivity of the identified neurons in the local bend network can account for our experimental results, and is suitable for reading out the neural population vector. Thus, for the first time to our knowledge, it is possible to link a proposed algorithm for neural population coding with synaptic and network mechanisms in an experimental system.

A light touch to the leech tubular body wall elicits a local bend. This stimulus activates two sets of mechanosensory neurons (P and T). The four P neurons encode the stimulus location and provide the major inputs to the local bend network^{7–9}, and the six T neurons encode the velocity of a stimulus¹⁰. The P neurons activate a layer of interneurons, that in turn activate a layer of 10 different motor neurons^{8,11} (Fig. 1a). The motor neurons control the length of longitudinal muscle and therefore determine body shape: shortening near the stimulus site and lengthening opposite the stimulus site produces a bend that is directed away from the stimulus.

We have recently quantified the responses (the number of spikes elicited) of the P neurons for different stimulus locations around the body perimeter⁹, and found that these neurons exhibit cosine-shaped tuning curves for stimulus location, with the tuning curve peaks (or preferred stimulus locations, P_i^*) distributed evenly around the body perimeter (Fig. 1b, c). Using a previously described neural decoding method¹², we estimated that stimulus location was encoded in the spike counts of the four P neurons with an error of 3% (root-mean-squared (r.m.s.) error expressed as a percentage of 360°). The local bend behaviour is directed to within 8% (r.m.s.) of stimulus location. The higher accuracy of the P neuron representation as compared with the behaviour suggests that a population

code based on the P neuron spike counts could be used by the local bend network.

Networks of neurons exhibiting cosine tuning curves are well suited for processing directional information, and are appropriately studied using neuronal decoding methods such as the population vector^{6,12}. The population vector is the sum of the preferred directions of a population of neurons, weighted by the respective spike counts, and can encode any stimulus location. In the case of P neurons, which exhibit cosine tuning and whose combined preferred locations form a two-dimensional cartesian coordinate system (see also ref. 13), the population vector is an optimal decoding method¹⁴. However, the existence of cosine tuning does not necessarily imply that the local bend network extracts information about touch location by computing a population vector from P neuron activity. For instance, touch location could be extracted from the P neuron that is most activated, using a 'winner-take-all' strategy¹⁵.

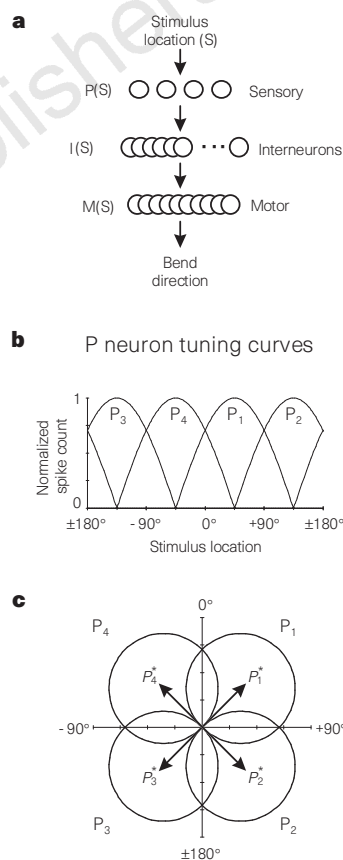


Figure 1 The local bend network and sensory tuning. **a**, Schematic of the local bend network that processes the location S of a touch stimulus to produce a bend directed away from the touch. The network is contained in a single segmental ganglion of the leech nervous system. There are three layers of neurons in the network. The number of neurons at sensory and motor levels is indicated by the number of circles. The exact number of interneurons is not known, although 17 have been identified so far⁸. **b**, The tuning properties of the mechanosensory P neurons to stimulus location. Shown are curves fitted to data described in Lewis⁹ and given by $P_i(S) = F(\cos(S - P_i^*))$ where S is the stimulus location, $P_i(S)$ is the normalized spike count, $F(x) = 0$ for $x \leq 0$ and $F(x) = x$ for $x > 0$. The perimeter of the cross-sectional body wall is roughly circular. We define a location $\theta = 0^\circ$ as the dorsal midline, $\theta = 180^\circ = -180^\circ$ as the ventral midline and $\theta = -90^\circ$ and $\theta = 90^\circ$ as the left and right lateral midlines, respectively. **c**, To emphasize the directionality of the neural responses and to allow comparison with the behavioural responses discussed later, the same P neuron tuning curves are expressed in a polar coordinate system. The P_i^* are the stimulus locations where the peak response occurs (that is, preferred locations). $P_i^* = (45^\circ, 135^\circ, -135^\circ, -45^\circ)$ for $i = (1, \dots, 4)$.

* Present address: Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, ON, Canada K1H 8M5.

To test the hypothesis that the local bend network extracts information encoded in a P neuron population vector, we recorded intracellular activity from P neurons along with behavioural output. We touched the skin at a location ($S = 45^\circ$) corresponding to the preferred location of the P_1 neuron; only P_1 was activated by this stimulus. The solid lines in Fig. 2a are unit vectors (directed towards the origin) indicating the bend direction for a single stimulus trial. Although the responses vary from trial to trial, the mean response (dotted line) is close to the stimulus location (7.9% r.m.s. error). In another set of trials, we injected current intracellularly to elicit spikes in the P_2 neuron (the previous touch stimulus did not activate P_2). The number of spikes elicited in P_2 was similar to the number produced in P_1 by the touch stimulus. This produced responses (Fig. 2b) that were on average close to the preferred location of P_2 (6.4% r.m.s. error). We then delivered both of these stimuli simultaneously (touch at $S = 45^\circ$, and P_2 current injection). The resulting responses were intermediate to and significantly different from those for each individual stimulus (Fig. 2c; $P < 0.05$). The effects of the two stimuli appear to be averaged and result in responses that were not statistically different from those produced by a single touch stimulus at $S = 90^\circ$ ($P > 0.05$). With the spike trains of both active P neurons, we calculated the P neuron population vector and found that it was significantly correlated with the bend direction ($r = 0.65$, $P < 0.05$). For comparison, the response to a touch stimulus at $S = 90^\circ$, which activates both P_1 and P_2 nearly equally, was similarly correlated with the P neuron population vector ($r = 0.55$, $P < 0.05$). These results eliminate the possibility that the local bend is governed by a winner-take-all mechanism, which predicts that the response to both stimuli would be the same as either of the responses to the individual stimuli, but not an intermediate response. In contrast, the results strongly suggest that touch location is computed using a P neuron population vector.

For the local bend network to extract and use information encoded in a population vector, the appropriate network and synaptic mechanisms must be present. Previous studies have partially characterized the local bend network^{8,11,16}. We reanalysed these data and found that the synaptic strength from P neurons to each of the 17 identified local bending interneurons was proportional to the cosine of the difference in their preferred stimulus locations⁹. This is exactly the pattern of connections that results in the accurate transfer of information encoded in a neural population vector¹⁷. Thus, the local bend network is appropriately wired to make use of a P neuron population vector.

In addition to cosine tuning and appropriate synaptic connec-

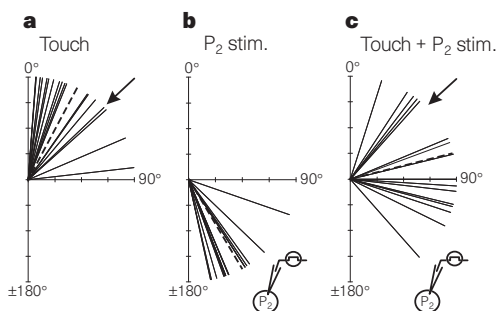


Figure 2 Bend directions elicited in different stimulus conditions in the experiments. In all panels, solid lines drawn in the polar coordinate system are unit vectors (directed toward the origin) corresponding to the bend direction. Dotted lines correspond to the mean bend direction. **a**, The responses to a touch stimulus at $S = 45^\circ$ (denoted by the arrow). **b**, The responses to intracellular current injection in the P_2 neuron (denoted by the icon). **c**, The responses elicited by simultaneous delivery of both stimuli. The data shown are from 15–16 trials in four animals.

tions, two questions remain regarding the accurate processing of the P neuron population vector: how many interneurons are required, and what are the effects of intrinsic neuronal variability? To address these issues, we used a neural network model based on the known features of the local bend network (see Methods). The network consisted of three layers: 4 P neurons, between 2 and 64 interneurons, and 10 motor neurons. We evaluated the accuracy of the model network by calculating the r.m.s. error between the stimulus location and the bend direction given by the model output, averaged over all stimulus locations. Figure 3a shows how the accuracy of the model depends on the number of interneurons as well as the noise parameter k (when $k = 0$ there is no noise; when $k = 1$ the standard deviation of the responses equals the mean). We do not know the variability (noise level) of a single interneuron or motor neuron; however, the variability in the electromyographic responses to identical stimuli corresponds to $k = 0.6$. The level of variability most associated with cortical neurons corresponds to $k = 1.0$, but can be as low as $k = 0.2$ under some conditions¹⁸. Also shown in Fig. 3a for comparison are experimental estimates of the local bend behavioural error (dotted line) and the P neuron decoding error (solid line)⁹. The accuracy of these model networks not only depends on the noise level (the parameter k) and the number of interneurons, but also on the distribution of the interneurons' preferred stimulus locations (I_i^*). For $k = 0$, the accuracy is limited only by the systematic error due to non-uniform distribution of I_i^* . As the number of interneurons increases beyond 10 (Fig. 3a), the networks are as accurate as they can be (that is, as accurate as the P neuron input). For even small noise levels however, networks consisting of four or fewer interneurons are less accurate than the actual behaviour. The minimum number of interneurons required for the model to be more accurate than the behaviour increases roughly linearly with noise level (N_i minimum $\approx 30k$, for $k \geq 0.2$; Fig. 3b). If the I_i^* are uniformly distributed, then four interneurons

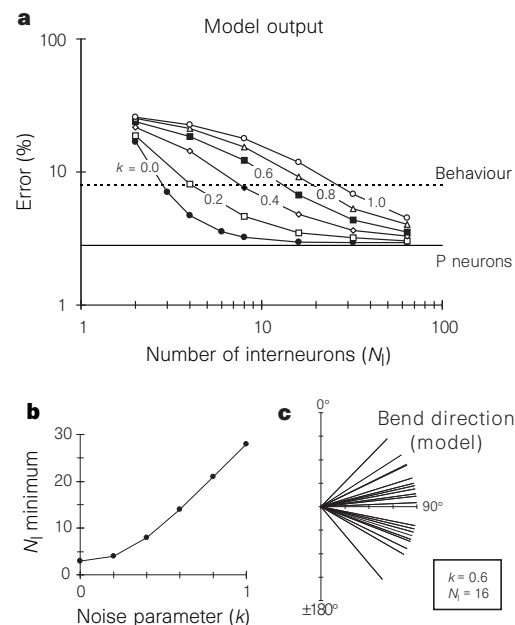


Figure 3 Results from the model. **a**, Summary of the errors produced by the model for different numbers of interneurons (N_i) and noise levels, k . Also shown for reference are the experimentally measured behavioural error (dotted line) and the P neuron decoding error (solid line). The filled circles show the results for zero noise level ($k = 0$). **b**, The minimum number of interneurons (N_i) required so that the model is more accurate than the experimentally observed behaviour. Minimum N_i increases almost linearly with the noise level. **c**, The responses of the model ($N_i = 16$) for 20 trials at a stimulus location of $S = 90^\circ$ (compare with Fig. 2c).

would be sufficient for the networks to achieve maximum accuracy (for $k = 0$). However, with increasing noise levels, more interneurons are required to average out this noise effectively.

At our best estimate of the actual noise level ($k = 0.6$), at least 14 interneurons are required to produce the experimentally observed behavioural accuracy. We analysed 10 randomly chosen networks consisting of 16 interneurons. In these networks, the error is due mostly to the noise level, k . The output of one of these networks, in response to a stimulus at $S = 90^\circ$, is shown in Fig. 3c. The model compares well with our experimental results (see Fig. 2c). For a stimulus at $S = 90^\circ$, the mean correlation coefficient between the P neuron population vector and the bend direction was 0.52. This suggests that the local bend network may only require 16 interneurons even though it is likely, considering the criteria used to find the 17 known interneurons, that there are 25 to 30 interneurons in the network⁸. One possibility is that we have underestimated the noise levels, and that more interneurons are required to compensate. Another possibility is that because some of these interneurons have roles in more than one behaviour¹⁹, a greater number of interneurons leads to more flexibility when choices are made between behaviours.

This study characterizes the population coding strategy of a neural system with a level of detail that has not been feasible in other systems. Although previous studies have correlated a neural population vector with behaviour², this is the first study that has done so on a trial-by-trial basis while monitoring and manipulating activity in the complete set of neurons that provides the basis for the population vector. With a minimum number of assumptions, a physiologically consistent model of the local bend network can account for the experimentally observed behaviour. It is very likely that this model embodies the key network mechanisms underlying the local bend behaviour. □

Methods

Experimental. We used previously described experimental procedures^{9,20}. The local bend reflex is segmental and can be studied in a reduced preparation consisting of a single innervated segment⁷. Mechanical stimuli (500 ms, 22 mN) were delivered using a nylon filament driven by a solenoid. We used standard intracellular recording techniques to monitor and manipulate spike activity in the P mechanosensory neurons. Spike trains were quantified by counting the number of spikes occurring in the 700 ms interval after stimulus onset. In the experiments outlined in Fig. 2, the P₂ neurons were electrically stimulated to produce a similar number of spikes as the average number of spikes elicited in P₁ by mechanical stimulation. To measure behavioural output, we recorded electromyographic activity at three locations (θ) along the body wall ($\theta = 45^\circ, 90^\circ$ and 135°). We fitted a cosine function to these data and the location of its peak provided an estimate of bend direction⁹. The behavioural error is given by the root-mean-squared (r.m.s.) difference between the mechanical stimulus location (S) and bend direction, expressed as a percentage of 360° . Statistical analyses were performed using standard methods for circular data²¹.

Neural population vector. We calculated the neural population vector for the P neuron spike counts as described previously^{6,9,12}. The direction of the population vector is given by the vector sum of the preferred locations (P_i^*) of the four P neurons, weighted by the respective spike counts.

Description of the model. A three-layered network was used to model the local bend network (Fig. 1a). The input layer consisted of four units (P_i) simulating the four P neurons. The output layer had 10 output units (M_n), representing the five bilateral classes of longitudinal motor neurons: dorsal, dorsolateral, lateral, ventrolateral, and ventral classes⁹. The middle layer of units (I_j) represented the local bending interneurons; we varied their number (N_I) in different simulations. The measure of neural activity can be thought of as normalized spike count (over the stimulus interval) for the P units and the post synaptic potential size for the I and M units. The properties of the P units were given by their experimentally derived tuning curves and preferred stimulus locations, P_i^* (Fig. 1b, c). The I and M units were characterized only by their preferred locations (I_j^* and M_n^*). The preferred locations of the identified interneurons and motor neurons, averaged over many animals, are evenly

distributed⁹, so we initially assign I_j^* and M_n^* to be uniformly distributed on the interval $(-180^\circ, 180^\circ)$. There is certainly variability between animals⁹ and presumably between different ganglia in the same animal, although this latter possibility has not been tested. In the model, we simulated the neuronal variability by adding a random number to each I_j^* and M_n^* . For the I_j^* , the range of the random number was set to 25% of the spacing between the initial I_j^* , corresponding to $\pm 0.25(360^\circ)/N_I$; for the M_n^* , it was fixed at $\pm 9^\circ$ (or $\pm 0.025(360^\circ)$). This results in variability similar to that found experimentally. Estimates of preferred location⁹ from experimental data vary, depending on the neuron, with an average range of $\pm 6^\circ$. The strength of the connectivity between units was proportional to the cosine of the difference between their respective preferred locations. The model network is summarized in equation (1), giving the activity levels for each P, I and M unit as a function of stimulus location (S):

$$\begin{aligned} P_i(S) &= \eta + \cos(S - P_i^*) \\ I_j(S) &= \eta + \sum_{i=1}^4 [w \cos(P_i^* - I_j^*)] F(P_i(S)) \\ M_n(S) &= \eta + \sum_{j=1}^{N_I} [w \cos(M_n^* - I_j^*)] F(I_j(S)) \end{aligned} \quad (1)$$

where $i = (1, \dots, 4)$, $j = (1, \dots, N_I)$, $n = (1, \dots, 10)$, η is a noise term, w is a scaling constant for the synaptic strength, and F is a threshold function defined as $F(x) = 0$ if $x \leq 0$ and $F(x) = x$ if $x > 0$. The ratio of post-synaptic to pre-synaptic voltage is about 0.6 for the physiological range of synaptic potentials²², so the parameter w was fixed at $w = 0.6$. Variability in a unit's activity was introduced by adding a gaussian random number η with zero mean, and a standard deviation such that the standard deviation σ of the response of a given unit with mean response, r , is $\sigma = kr$ (k = coefficient of variation). The same value of k was used for both I and M units. For a given stimulus location S , the activity of each P unit was a random number drawn from a normal distribution with mean p given by the P neurons' tuning curves, and the experimentally measured⁹ standard deviation σ_p given by the equation $\sigma_p = 0.08 + 0.17p$.

Simulation procedure. A network was defined by choosing the number of I units (N_I), the preferred locations for the I and M units, (I_j^* and M_n^*), and the value of the noise parameter (k). The input to the network was a given stimulus location S . The resulting output of the model was the set of activity levels in the M units. To obtain an estimate of the bend direction, we used the population vector formed by the sum of the M_n^* weighted by their respective activity levels. The model error was the difference between the stimulus location and the estimated bend direction (expressed as a percentage of 360°). This procedure was repeated for 5,000 trials, with a different random choice of I_j^* , M_n^* , and S in each trial, to obtain the r.m.s. error for a network of a size given by N_I and noise level k .

Received 18 August; accepted 7 October 1997.

1. Sparks, D. L., Kristan, W. B. Jr & Shaw, B. K. in *Neurons, Networks, and Motor Behavior* (eds Stein, P. S. G., Grillner, S., Selverston, A. I. & Stuart, D. G.) 21–32 (MIT Press, Cambridge, MA, 1997).
2. Georgopoulos, A. New concepts in generation of movement. *Neuron* **13**, 257–268 (1994).
3. Lee, C., Rohrer, W. H. & Sparks, D. L. Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature* **332**, 357–360 (1988).
4. Groh, J., Born, R. & Newsome, W. How is a sensory map read out? Effects of microstimulation in visual area MT on saccades and smooth pursuit eye movements. *J. Neurosci.* **17**, 4312–4330 (1997).
5. Kristan, W. B. Jr, McGirr, S. J. & Simpson, G. V. Behavioural and mechanosensory neurone responses to skin stimulation in leeches. *J. Exp. Biol.* **96**, 143–160 (1982).
6. Georgopoulos, A. P., Schwartz, A. B. & Kettner, R. E. Neuronal population coding of movement direction. *Science* **233**, 1416–1419 (1986).
7. Kristan, W. B. Jr Sensory and motor neurones responsible for the local bending response in leeches. *J. Exp. Biol.* **96**, 161–180 (1982).
8. Lockery, S. R. & Kristan, W. B. Jr Distributed processing of sensory information in the leech. II. Identification of interneurons contributing to the local bending reflex. *J. Neurosci.* **10**, 1816–1829 (1990).
9. Lewis, J. E. *From Touch Localization to Directed Behavior: Neural Computation in the Leech Thesis* (Univ. California, San Diego, San Diego, CA, 1997).
10. Carlton, T. & McVean, A. The role of touch, pressure and nociceptive mechanoreceptors of the leech in unrestrained behaviour. *J. Comp. Physiol. A* **177**, 781–791 (1995).
11. Lockery, S. R. & Kristan, W. B. Jr Distributed processing of sensory information in the leech. I. Input-output relations of the local bending reflex. *J. Neurosci.* **10**, 1811–1815 (1990).
12. Salinas, E. & Abbott, L. F. Vector reconstruction from firing rates. *J. Comp. Neurosci.* **1**, 89–107 (1994).
13. Miller, J. P., Jacobs, G. A. & Theunissen, F. E. Representation of sensory information in the cricket cercal sensory system. I. Response properties of the primary interneurons. *J. Neurophysiol.* **66**, 1680–1689 (1991).
14. Abbott, L. F. Decoding neuronal firing and modeling neural networks. *Q. Rev. Biophys.* **27**, 291–331 (1994).
15. Salzman, C. D. & Newsome, W. T. Neural mechanisms for forming a perceptual decision. *Science* **264**, 231–237 (1994).

16. Lockery, S. R., Wittenberg, G., Kristan, W. B. Jr & Cottrell, G. W. Function of identified interneurons in the leech elucidated using neural networks trained by back-propagation. *Nature* **340**, 468–471 (1989).
17. Salinas, E. & Abbott, L. F. Transfer of coded information from sensory to motor networks. *J. Neurosci.* **15**, 6461–6474 (1995).
18. Gur, M., Beylin, A. & Snodderly, D. M. Response variability of neurons in primary visual cortex (V1) of alert monkeys. *J. Neurosci.* **17**, 2914–2920 (1997).
19. Shaw, B. K. & Kristan, W. B. Jr The neuronal basis of the behavioral choice between swimming and shortening in the leech: control is not selectively exercised at higher circuit levels. *J. Neurosci.* **17**, 786–795 (1997).
20. Muller, K. J., Nicholls, J. G. & Stent, G. S. *Neurobiology of the Leech* (Cold Spring Harbor Laboratory Press, New York, 1981).
21. Batschelet, E. *Circular Statistics in Biology* (Academic, New York, 1981).
22. Lockery, S. R. & Sejnowski, T. J. Distributed processing of sensory information in the leech. III. A dynamical neural network model of the local bending reflex. *J. Neurosci.* **12**, 3877–3895 (1992).

Acknowledgements. We thank M. Courtemanche, S. Kealey, T. Lewis and K. Moortgat for comments on the manuscript. This work was supported by a research grant from NIMH to W.B.K. and a NRSA graduate fellowship from NIMH to J.E.L.

Correspondence and requests for materials should be addressed to W.B.K. (e-mail: wkristan@ucsd.edu).

The peroxisome proliferator-activated receptor- γ is a negative regulator of macrophage activation

Mercedes Ricote^{†,‡}, Andrew C. Li^{†,‡}, Timothy M. Willson[§], Carolyn J. Kelly[¶] & Christopher K. Glass^{*}

^{*} Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0651, USA

[†] Divisions of Cellular and Molecular Medicine, and Endocrinology and Metabolism, [‡] Division of Cardiology, and [¶] Division of Nephrology and the San Diego Veterans Affairs Medical Center, La Jolla, California 92161, USA
[§] Department of Medicinal Chemistry, Glaxo Wellcome Research and Development, PO Box 13398, Research Triangle Park, North Carolina 27709, USA

The peroxisome proliferator-activated receptor- γ (PPAR- γ) is a member of the nuclear receptor superfamily of ligand-dependent transcription factors that is predominantly expressed in adipose tissue, adrenal gland and spleen^{1–3}. PPAR- γ has been demonstrated to regulate adipocyte differentiation and glucose homeostasis in response to several structurally distinct compounds, including thiazolidinediones and fibrates^{3–6}. Naturally occurring compounds such as fatty acids and the prostaglandin D₂ metabolite 15-deoxy- $\Delta^{12,14}$ prostaglandin J₂ (15d-PGJ₂) bind to PPAR- γ and stimulate transcription of target genes^{7–10}. Prostaglandin D₂ metabolites have not yet been identified in adipose tissue, but are major products of arachidonic-acid metabolism in macrophages¹¹, raising the possibility that they might serve as endogenous PPAR- γ ligands in this cell type. Here we show that PPAR- γ is markedly upregulated in activated macrophages and inhibits the expression of the inducible nitric oxide synthase, gelatinase B and scavenger receptor A genes in response to 15d-PGJ₂ and synthetic PPAR- γ ligands. PPAR- γ inhibits gene expression in part by antagonizing the activities of the transcription factors AP-1, STAT and NF- κ B. These observations suggest that PPAR- γ and locally produced prostaglandin D₂ metabolites are involved in the regulation of inflammatory responses, and raise the possibility that synthetic PPAR- γ ligands may be of therapeutic value in human diseases such as atherosclerosis and rheumatoid arthritis in which activated macrophages exert pathogenic effects.

To determine whether 15d-PGJ₂ might be involved in the regulation of macrophage gene expression by activating PPAR- γ , we performed RNase protection assays using RNA derived from different macrophage populations. Resting bone-marrow-derived macrophages express low levels of PPAR- γ mRNA, whereas acti-

vated peritoneal macrophages express high levels of PPAR- γ (Fig. 1a). Peritoneal macrophages treated with 15d-PGJ₂ exhibited morphological features typical of resting cells even in the presence of interferon- γ (IFN- γ), which is a potent inducer of macrophage activation (Fig. 1b). Similar effects were observed in the presence of BRL 49653 (Fig. 1b), which is a specific PPAR- γ agonist of the thiazolidinedione family⁵ (Fig. 1b). These observations suggest that PPAR- γ might inhibit the expression of genes that become up-regulated during macrophage differentiation and activation. To evaluate this possibility, we examined the effects of 15d-PGJ₂ and

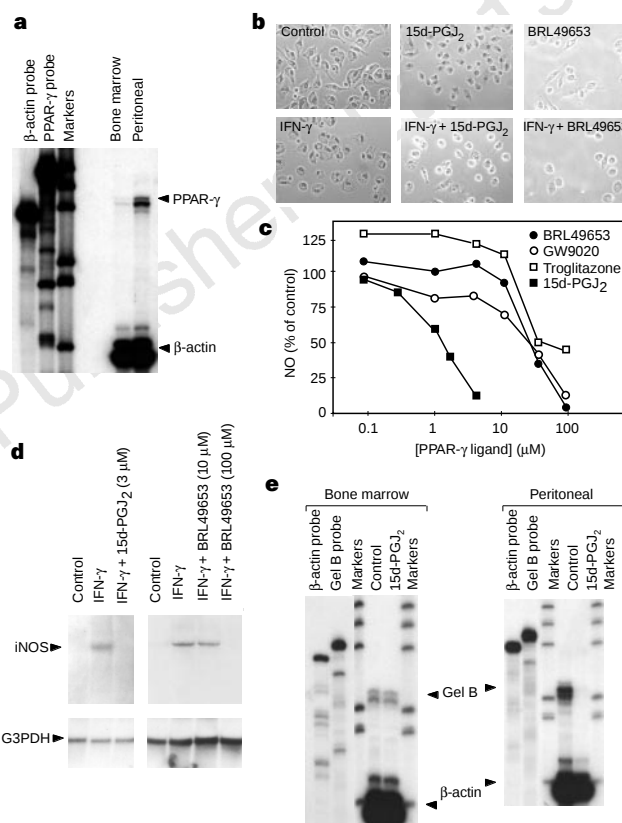


Figure 1 Macrophage expression of PPAR- γ and inhibition of iNOS and gelatinase B gene expression by 15d-PGJ₂. **a**, PPAR- γ mRNA is highly expressed in thioglycolate-elicited peritoneal macrophages as determined by RNase protection assay, but not in resting macrophages derived from bone-marrow progenitor cells in the presence of M-CSF. Positions of the protected mRNA fragments for PPAR- γ and β -actin are indicated. **b**, 15d-PGJ₂ and synthetic PPAR- γ ligands inhibit morphological changes in macrophages induced by IFN- γ . Mouse peritoneal macrophages were plated in plastic chamber slides, treated with control solvent, IFN- γ (500 U ml⁻¹), and 15d-PGJ₂ or BRL 49653 for 18 h as indicated, and photographed using phase-contrast microscopy. **c**, 15d-PGJ₂ inhibits NO production in response to IFN- γ treatment. Thioglycolate-elicited peritoneal macrophages were treated with IFN- γ (500 U ml⁻¹) and the indicated concentrations of 15d-PGJ₂ or synthetic PPAR- γ ligands for 24 h before measurement of nitrite concentrations in the culture supernatant. **d**, 15d-PGJ₂ and BRL 49653 inhibit induction of iNOS mRNA following IFN- γ treatment. Mouse peritoneal macrophages were incubated with control solvent, IFN- γ (500 U ml⁻¹), and/or 15d-PGJ₂ or BRL 49653 for 18 h at the indicated concentrations. Total RNA was isolated and subjected to northern blot analysis using a specific iNOS cDNA probe. **e**, Regulation of gelatinase B mRNA levels by 15d-PGJ₂ in mouse peritoneal macrophages but not bone-marrow-derived macrophages. Mouse bone-marrow progenitor cells were cultured for 3 days in M-CSF to induce macrophage differentiation in the presence or absence of 3 μ M 15d-PGJ₂. Similarly, thioglycolate-elicited mouse peritoneal macrophages were cultured for 18 h in the presence or absence of 2 μ M 15d-PGJ₂. Total RNA was isolated and subjected to analysis by RNase protection assay using specific antisense cRNAs for gelatinase B and β -actin, with protected mRNA fragments indicated.

other PPAR- γ ligands on expression of the inducible nitric oxide synthase (iNOS), which exerts cytotoxic effects on invading microorganisms and is upregulated in activated macrophages in response to IFN- γ ¹². Treatment of peritoneal macrophages with increasing concentrations of 15d-PGJ₂ or synthetic PPAR- γ ligands led to a dose-dependent inhibition of IFN- γ -dependent nitrite production (Fig. 1c) and inhibited the induction of iNOS mRNA (Fig. 1d). Although highly specific for PPAR- γ ^{5,6,9,13}, BRL 49653, troglitazone and GW 2090 were less potent than the natural ligand 15d-PGJ₂ as inhibitors of iNOS activity.

As a second marker of macrophage gene expression, we examined effects of 15d-PGJ₂ on expression of gelatinase B, a matrix metalloproteinase that is upregulated during macrophage activation and contributes to tissue damage in acute and chronic inflammation^{14,15}. Treatment of bone-marrow-derived macrophages that express low levels of PPAR- γ with 15d-PGJ₂ had little effect on gelatinase B mRNA levels (Fig. 1e). In contrast, 15d-PGJ₂ treatment of peritoneal macrophages that express high levels of PPAR- γ resulted in a near-complete inhibition of gelatinase B mRNA expression (Fig. 1e). To further investigate the mechanisms of inhibition, the effects of 15d-PGJ₂ on regulation of the gelatinase B promoter were studied in U937 histiocytic leukaemia cells. These cells express PPAR- γ ¹⁶, and can be induced to differentiate into macrophage-like cells by treatment with the phorbol ester TPA. When U937 cells were transfected with a gelatinase B promoter-luciferase reporter gene, TPA treatment resulted in a marked increase in promoter activity

(Fig. 2a). This activity was strongly inhibited by concurrent treatment of the cells with 15d-PGJ₂. Overexpression of PPAR- γ potentiated the inhibitory effects of 15d-PGJ₂, suggesting that they were mediated by PPAR- γ (Fig. 2a). Transcriptional activation of the gelatinase B promoter is dependent on binding sites for AP-1 proteins present in proximal and distal regulatory regions^{17,18} (Fig. 2a). To determine whether other AP-1-dependent promoters are also subject to negative regulation by 15d-PGJ₂, we examined its effects on transcriptional activation of the scavenger receptor A (SR-A) gene. The SR-A gene encodes a macrophage-specific cell-surface protein implicated in Ca²⁺-independent cell-adhesion events and the uptake of polyanionic macromolecules, including oxidized low-density lipoprotein¹⁹. Transcriptional activation of the SR-A promoter in response to TPA is dependent on combinatorial interactions between AP-1 and Ets transcription factors²⁰. Similar to the results observed for gelatinase B, 15d-PGJ₂ inhibited TPA-dependent activation of the SR-A promoter in a manner that was potentiated by overexpression of PPAR- γ (Fig. 2b).

To examine the mechanisms by which 15d-PGJ₂ inhibits iNOS expression, we investigated regulation of the iNOS promoter in RAW 264.7 macrophages. As well as having binding sites for AP-1 proteins, the iNOS promoter also contains multiple binding sites for Stat1 and NF- κ B (Fig. 2d), which are required for maximal responses to IFN- γ and lipopolysaccharide (LPS), respectively^{21,22}. RAW 264.7 cells express very low levels of PPAR- γ mRNA (data not shown), such that activation of a PPAR-dependent promoter (AOx-TK)

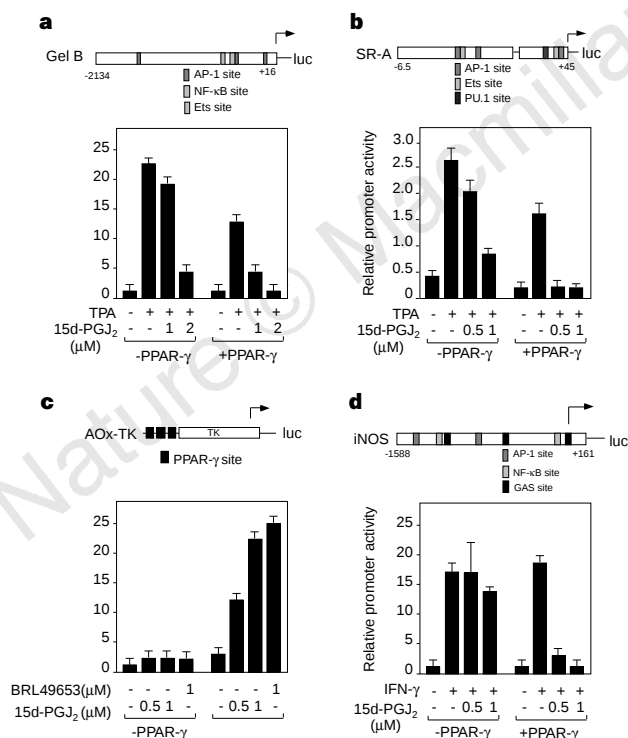


Figure 2 15d-PGJ₂ inhibits the gelatinase B, SR-A and iNOS promoters in a PPAR- γ -dependent manner. **a**, U937 cells transfected with a luciferase reporter plasmid (luc) under transcriptional control of the gelatinase B (Gel B) promoter. **b**, U937 cells transfected with a luciferase reporter plasmid under transcriptional control of SR-A regulatory elements. **c**, RAW 264.7 cells transfected with a PPAR-responsive reporter gene containing three copies of the acyl CoA oxidase PPAR response element linked to the TK promoter (AOx-TK). **d**, RAW 264.7 cells transfected with a iNOS promoter-luciferase reporter plasmid. Cells were co-transfected with 1 μ g of reporter plasmid and either 100 ng of a PPAR- γ expression plasmid or control DNA as indicated. Cells were treated with combinations of 15d-PGJ₂, TPA (0.1 μ M) and/or IFN- γ (100 U ml⁻¹) as shown and collected for analysis of reporter gene activity 24 h later.

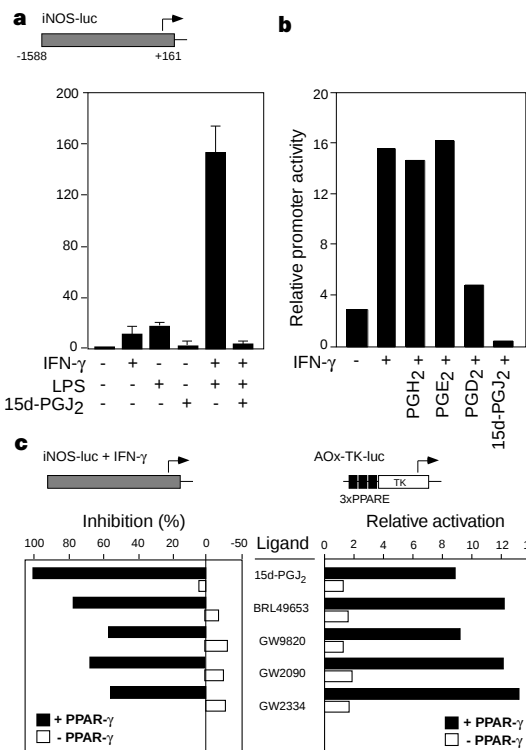


Figure 3 iNOS promoter activity is inhibited by prostaglandin D₂ metabolites and synthetic PPAR- γ ligands. **a**, 15d-PGJ₂ inhibits the synergistic transcriptional response of the iNOS promoter to the combination of LPS and IFN- γ . RAW 264.7 cells were transfected with a luciferase reporter gene under transcriptional control of the iNOS promoter and a PPAR- γ expression plasmid. Cells were treated with combinations of LPS (5 μ g ml⁻¹), IFN- γ (100 U ml⁻¹) and 15d-PGJ₂ and analysed for luciferase activity 24 h later. **b**, RAW 264.7 cells were transfected with the iNOS reporter gene and a PPAR- γ expression plasmid and treated with IFN- γ and the indicated prostaglandins. Cells were analysed for luciferase activity 24 h later. **c**, Cells were transfected with a PPAR- γ expression plasmid (+PPAR- γ) or control DNA (-PPAR- γ) and luciferase reporter genes under the control of either the iNOS promoter (left) or the AOx-TK promoter (right). Cells were treated with 1 μ M 15d-PGJ₂ or the indicated synthetic PPAR- γ ligands and analysed 24 h later.

required transfection of a PPAR- γ expression plasmid (Fig. 2c). This cell line therefore allowed a direct assessment of the role of PPAR- γ in mediating the inhibitory effects of 15d-PGJ₂ and synthetic PPAR- γ ligands on macrophage gene expression. In the absence of a co-transfected PPAR- γ expression plasmid, treatment of RAW 264.7 macrophages with 15d-PGJ₂ at concentrations up to 1 μ M had little or no effect on iNOS activation by IFN- γ (Fig. 2d). However, when a PPAR- γ expression plasmid was transfected into these cells, 15d-PGJ₂ strongly inhibited IFN- γ -dependent activation of the iNOS promoter (Fig. 2d), as well as the synergistic transcriptional response to the combination of IFN- γ and LPS (Fig. 3a). Thus 15d-PGJ₂ inhibits induction of the iNOS promoter by a PPAR- γ -dependent mechanism. The inhibitory effects of 15d-PGJ₂ and PPAR- γ were promoter specific, because no effect was observed on the expression of several other viral and cellular promoters, including the β -actin, HSV thymidine kinase, RSV and CMV promoters (data not shown).

Inhibition of IFN- γ -dependent induction of iNOS promoter activity in RAW 264.7 cells was observed for the 15d-PGJ₂ precursor, PGD₂, but not for other prostaglandin metabolites, including PGE₂ and PGH₂ (Fig. 3b). Because PGD₂ does not bind to PPAR- γ ^{7,8}, and no inhibitory activity was observed in the absence of co-transfected PPAR- γ (data not shown), these results indicate metabolic conversion of PGD₂ to an activating ligand by RAW 264.7 macrophages. PPAR- γ -dependent inhibition of the iNOS promoter was also observed for several synthetic ligands (Fig. 3c), including the thiazolidinedione BRL 49653, as well as the fibric-acid derivatives GW9820, GW2090 and GW2334 (Fig. 3c).

Because activation of the SR-A, gelatinase B and iNOS promoters depends on combinatorial interactions between members of the AP-1, STAT and NF- κ B families of transcription factors, we evaluated minimal promoters containing binding sites for these factors to determine whether they might be targets for negative regulation

by 15d-PGJ₂ and synthetic PPAR- γ ligands. We found that 15d-PGJ₂ effectively inhibited each of these promoters in a PPAR- γ -dependent manner (Fig. 4), but did not inhibit a promoter containing binding sites for SP-1 (data not shown). Thus activation of PPAR- γ by 15d-PGJ₂ results in inhibitory effects on at least three different classes of transcription factors that play general roles in regulating inflammatory responses in many cell types, including macrophages. The synthetic PPAR- γ ligand BRL 49653 also inhibited STAT, NF- κ B and AP-1 activities in a PPAR- γ -dependent manner, but did so less effectively than 15d-PGJ₂, particularly in the case of the AP-1-dependent promoter. These results are consistent with the requirement for higher concentrations of BRL 49653 to inhibit iNOS expression in primary macrophages and RAW 264.7 cells transfected with PPAR- γ expression vectors.

Previous studies have demonstrated that prostaglandin D₂ synthase, which is required for 15d-PGJ₂ synthesis, is predominantly expressed in macrophages and specialized antigen-presenting cells¹¹. The receptor PPAR- α , which is related to PPAR- γ , can be weakly activated by 15d-PGJ₂, but PPAR- α is not expressed in activated macrophages (presented as Supplementary information). Although activated macrophages express both PPAR- γ (Fig. 1) and PPAR- δ (data not shown), 15d-PGJ₂ is not an effective ligand for PPAR- δ ¹⁰. These observations therefore raise the possibility that, in addition to its proposed roles in the regulation of adipocyte differentiation and glucose metabolism, PPAR- γ may also regulate inflammatory responses as a result of local production of 15d-PGJ₂. Taken together with recent evidence that fatty acids are physiological ligands of PPAR- γ in adipose tissue⁹, our results suggest that PPAR- γ is regulated by distinct classes of ligands in different locations. PPAR- γ is under the transcriptional control of at least two distinct promoters³, suggesting that different signals regulate its expression in fat cells and macrophages. With regard to the different efficacies of natural and synthetic PPAR- γ ligands as inhibitors of macrophage gene expression, we note that dissociation of target gene activation and AP-1 inhibition has previously been described for ligands of the retinoic acid receptor²³. Thiazolidinediones, 15d-PGJ₂ and fibric acid derivatives represent three structurally distinct classes of activating ligands, and are thus likely to induce different receptor conformations that account for these observations. Taken together, these results suggest that synthetic PPAR- γ ligands that are analogues of 15d-PGJ₂ may have therapeutic applications in diseases in which activated macrophages play prominent pathogenic roles, such as rheumatoid arthritis and atherosclerosis.

Methods

Cell culture. Murine bone-marrow progenitor cells and primary peritoneal macrophages were isolated and cultured as described²⁴. Recombinant macrophage colony-stimulating factor (M-CSF) was used at 20 ng ml⁻¹. Murine IFN- γ was used at 500 U ml⁻¹. U-937 (ATCC) and THP-1 (ATCC) cells were cultured in RPMI 1640 supplemented with 10% heat-inactivated fetal calf serum (Gemini), 100 U ml⁻¹ penicillin and 100 mg ml⁻¹ streptomycin. RAW 264.7 (ATCC) cells and HeLa (ATCC) cells were cultured under the same conditions except that Dulbecco's modified Eagle's medium (DMEM) was used.

RNA analysis. Total RNA was isolated by the guanidium thiocyanate method and RNase protection assays were performed as described²⁵. The antisense probe for gelatinase B corresponded to nucleotides 1779–1999. The antisense PPAR- γ probe corresponded to nucleotides 800–1093. Northern blot analysis of iNOS mRNA used 10 μ g of total RNA per treatment conditions. The iNOS probe consisted of a 500-base pair fragment of the cDNA corresponding to nucleotides 2849–3349. To control for equivalency of RNA loading and transfer, blots were hybridized with the cDNA for glyceraldehyde-3-phosphate dehydrogenase. Probes were labelled by random priming.

Nitrite concentrations. Supernatants from thioglycolate-elicited peritoneal macrophages cultured with or without IFN- γ in the presence or absence of 15d-PGJ₂ or synthetic ligands were mixed with an equal volume of the Griess

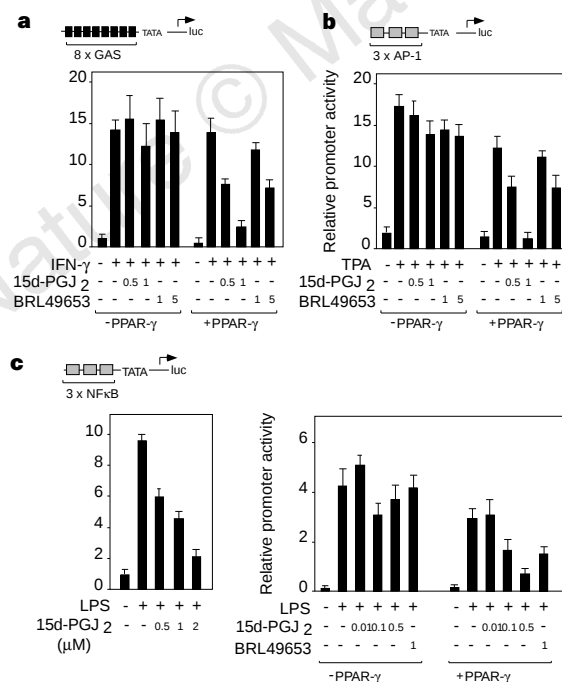


Figure 4 15d-PGJ₂ inhibits transcriptional responses mediated by AP-1, Stat1 and NF- κ B transcription factors. Cells were transfected with a PPAR- γ expression vector or control DNA and reporter genes consisting of a minimal TATA-containing promoter linked to multiple GAS elements that are binding sites for Stat1 (a), AP-1 (b) or NF- κ B (c). Cells were treated with the indicated concentrations (in μ M) of 15d-PGJ₂ or BRL 49653 and TPA (0.1 μ M), IFN- β (100 U ml⁻¹) or LPS (5 μ g ml⁻¹) and analysed for luciferase activity 24 h later. HeLa cells, which express low levels of PPAR- γ , were used in the experiments shown in a and b.

reagent in 96-well plates, gently shaken for 20 min at room temperature, and read in a microplate reader at 550 nm (ref. 26). Readings for experimental samples were read off a standard curve generated by dilutions of NaNO₂. The assay is sensitive to nitrite concentrations of approximately 1 μM.

Transcription assays. U937 cells were transiently transfected with luciferase reporter genes by electroporation²⁷ in 200 μl OptiMEM (Gibco) with 10 μg reporter plasmid and 0.1 μg β-actin-lacZ reporter plasmid, as internal transfection control, at 250 V, 960 μF using a Biorad electroporator with capacitance extender. RAW 267.4 cells were transfected using lipofectamin (Gibco-BRL). HeLa cells were transfected using Ca²⁺ phosphate. Salmon sperm DNA or vector DNA was used to balance all transfection groups to equal amounts of transfected DNA. Luciferase and β-galactosidase enzymatic activities were determined²⁰ and luciferase activity was normalized to the β-galactosidase standard. The reporter gene constructs for the SR-A gene, NOS, gelatinase B, AOX-TK and β-actin have been described^{18,20,22}.

Received 11 September; accepted 27 October 1997.

1. Kliewer, S. A. *et al.* Differential expression and activation of a family of murine peroxisome proliferator-activated receptors. *Proc. Natl Acad. Sci. USA* **91**, 7355–7359 (1994).
2. Lemberger, T., Desvergne, B. & Wahli, W. PPARs: a nuclear receptor signaling pathway in lipid metabolism. *Annu. Rev. Cell Dev. Biol.* **12**, 335–363 (1996).
3. Tontonoz, P., Hu, E., Graves, R. A., Budavari, A. I. & Spiegelman, B. M. mPPAR gamma 2: tissue-specific regulator of an adipocyte enhancer. *Genes Dev.* **8**, 1224–1234 (1994).
4. Tontonoz, P., Hu, E. & Spiegelman, B. M. Stimulation of adipogenesis in fibroblasts by PPARγ2, a lipid-activated transcription factor. *Cell* **79**, 1147–1156 (1994).
5. Lehmann, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor gamma (PPARγ). *J. Biol. Chem.* **270**, 12953–12956 (1995).
6. Willson, T. M. *et al.* The structure-activity relationship between peroxisome proliferator-activated receptor gamma agonism and the antihyperglycemic activity of thiazolidinediones. *J. Med. Chem.* **39**, 665–668 (1996).
7. Kliewer, S. A. *et al.* A prostaglandin J₂ metabolite binds peroxisome proliferator-activated receptor and promotes adipocyte differentiation. *Cell* **83**, 813–819 (1995).
8. Forman, B. M. *et al.* 15-Deoxy-Δ^{12,14}-prostaglandin J₂ is a ligand for the adipocyte determination factor PPARγ. *Cell* **83**, 803–812 (1995).
9. Kliewer, S. A. *et al.* Fatty acids and eicosanoids regulate gene expression through direct interactions with peroxisome proliferator-activated receptors α and γ. *Proc. Natl Acad. Sci. USA* **94**, 4318–4323 (1997).
10. Krey, G. *et al.* Fatty acids, eicosanoids, and hypolipidemic agents identified as ligands of peroxisome proliferator-activated receptors by coactivator-dependent receptor ligand assay. *Mol. Endocrinol.* **11**, 779–791 (1997).
11. Urade, Y., Ujihara, M., Horiguchi, Y., Ikai, K. & Hayaishi, O. The major source of endogenous prostaglandin D₂ production is likely antigen-presenting cells. *J. Immunol.* **143**, 2982–2989 (1989).
12. Lowenstein, C. J. *et al.* Macrophage nitric oxide synthase gene: two upstream regions mediate induction by interferon gamma and lipopolysaccharide. *Proc. Natl Acad. Sci. USA* **90**, 8730–8734 (1993).
13. Hawke, R. L. *et al.* Potent hypocholesterolemic activity of novel ureido phenoxysobutyrate correlates with their intrinsic fibrinolytic potency and not with their ACAT inhibitory activity. *J. Lipid Res.* **38**, 1189–1203 (1997).
14. Matrisian, L. M. The matrix degrading metalloproteinases. *BioEssays* **14**, 455–463 (1992).
15. Welgus, H. G., Campbell, E. J. & Cury, J. D. Neutral metalloproteinases produced by human mononuclear phagocytes. *J. Clin. Invest.* **86**, 1496–1502 (1990).
16. Greene, M. E. *et al.* Isolation of the human peroxisome proliferator activated receptor gamma cDNA: expression in hematopoietic cells and chromosomal mapping. *Gene Express.* **4**, 281–299 (1995).
17. Saarialho-Kere, U. K., Welgus, H. G. & Parks, W. C. Distinct mechanisms regulate interstitial collagenase and 92-kDa gelatinase expression in human monocytic-like cells exposed to bacterial endotoxin. *J. Biol. Chem.* **268**, 17354–17361 (1993).
18. Sato, H. & Seiki, M. Regulatory mechanism of 92 kDa type IV collagenase gene expression which is associated with invasiveness of tumor cells. *Oncogene* **8**, 395–405 (1993).
19. Krieger, M. & Herz, J. Structures and functions of multiligand lipoprotein receptors: macrophage scavenger receptors and LDL receptor-related protein (LRP). *Annu. Rev. Biochem.* **63**, 601–637 (1994).
20. Wu, H., Moulton, K., Horvai, A., Parik, S. & Glass, C. K. Combinatorial interactions between AP-1 and ets-domain proteins contribute to the developmental regulation of the macrophage scavenger gene. *Mol. Cell. Biol.* **14**, 2129–2139 (1994).
21. Xie, Q. W., Whisnant, R. & Nathan, C. Promoter of the mouse gene encoding calcium-independent nitric oxide synthase confers inducibility by interferon gamma and bacterial lipopolysaccharide. *J. Exp. Med.* **177**, 1779–1784 (1993).
22. Lorschach, R. B., Murphy, W. J., Lowenstein, C. J., Snyder, S. H. & Russell, S. W. Expression of the nitric oxide synthase gene in mouse macrophages activated for tumor cell killing. Molecular basis for the synergy between interferon-gamma and lipopolysaccharide. *J. Biol. Chem.* **268**, 1908–1913 (1993).
23. Fanjul, A. *et al.* A new class of retinoids with selective inhibition of AP-1 inhibits proliferation. *Nature* **372**, 107–111 (1994).
24. Horvai, A. E. *et al.* Nuclear integration of JAK/STAT and Ras/AP-1 signaling by CBP and p300. *Proc. Natl Acad. Sci. USA* **94**, 1074–1079 (1997).
25. Ausubel, F. *et al.* *Current Protocols in Molecular Biology* 1-0 (John Wiley, New York, 1990).
26. Green, L. C. *et al.* Analysis of nitrate, nitrite, and [¹⁵N]nitrate in biological fluids. *Biochemistry* **126**, 131–138 (1982).
27. Zhu, Y., Pless, M., Inhorn, R., Mathey-Prevot, B. & D'Andrea, A. D. The murine *DUB-1* gene is specifically induced by the βc subunit of the interleukin-3 receptor. *Mol. Cell. Biol.* **16**, 4808–4817 (1996).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank C. Archer for assistance with nitrate measurements, and T. Schneiderman for assistance with manuscript preparation. A.C.L. is supported by a Physician Scientist grant from the NIH. C.J.K. is a clinical investigator of the Medical Research Service, Department of Veterans Affairs. C.K.G. is an established investigator of the American Heart Association. These studies were supported by NIH grants to C.J.K. and C.K.G.

Correspondence and requests for materials should be addressed to C.K.G. (e-mail: cglass@ucsd.edu).

PPAR-γ agonists inhibit production of monocyte inflammatory cytokines

Chengyu Jiang, Adrian T. Ting & Brian Seed

Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

The peroxisome proliferator-activated receptor-γ (PPAR-γ) is a member of the nuclear receptor family of transcription factors, a large and diverse group of proteins that mediate ligand-dependent transcriptional activation and repression^{1,2}. Expression of PPAR-γ is an early and pivotal event in the differentiation of adipocytes^{3–6}. Several agents that promote differentiation of fibroblast lines into adipocytes have been shown to be PPAR-γ agonists^{7–10}, including several prostanoids, of which 15-deoxy-Δ^{12,14}-prostaglandin J₂ is the most potent^{8,9}, as well as members of a new class of oral antidiabetic agents, the thiazolidinediones⁷, and a variety of non-steroidal anti-inflammatory drugs (NSAIDs)¹⁰. Here we show that PPAR-γ agonists suppress monocyte elaboration of inflammatory cytokines at agonist concentrations similar to those found to be effective for the promotion of adipogenesis. Inhibition of cytokine production may help to explain the incremental therapeutic benefit of NSAIDs observed in the treatment of rheumatoid arthritis at plasma drug concentrations substantially higher than are required to inhibit prostaglandin G/H synthase (cyclooxygenase).

In adipose tissue the adipogenic actions of PPAR-γ agonists are opposed by several cytokines, including tumor necrosis factor-α, (TNF-α)¹¹, interferon-γ¹², interleukin (IL)-1, IL-6, transforming growth factor-β¹³, and leukaemia inhibitory factor^{14,15}, which all inhibit adipogenesis or lipolysis. Systemic infusion of TNF-α in rats causes insulin resistance¹⁶, and in culture the antidiabetic thiazolidinediones block inhibition by TNF-α of both adipogenesis and insulin sensitivity¹⁷. PPAR-γ is also expressed in the immune system, in the spleen white and red pulp of rats¹⁸, and by human monocytes and bone-marrow precursors¹⁹. Although leukotriene B₄, an agonist of the related receptor PPAR-α, has anti-inflammatory activity²⁰, the role of PPAR-γ ligands in inflammation has not been studied so far.

To explore this, we prepared human peripheral blood mononuclear cells by plastic adherence from whole-blood buffy-coat fractions, and induced cytokine synthesis using phorbol myristyl acetate (PMA). The effect of prostaglandins and related agents on cytokine production was evaluated by adding the agents at the same time as the inducer. Cytokine synthesis induced by PMA can be inhibited by prostanoids of the prostaglandin J₂ family, whereas other prostaglandins either have relatively little effect or potentiate cytokine production (Fig. 1). A similar pattern of activity has been reported for prostaglandins that are known to act at least in part through activation of PPAR-γ. Consistent with this observation, the structurally dissimilar PPAR-γ agonist troglitazone inhibited PMA-induced TNF-α synthesis, whereas agents acting through the related PPAR-α receptor (8[S]-hydroxyeicosatetraenoic acid, eicosatetraenoic acid, WY-14643, or leukotriene B₄)^{20,21}, the thyroid hormone receptor, or retinoid receptors, had no activity (Fig. 1).

Cells of the monocyte-macrophage lineage can be induced to secrete inflammatory cytokines by exposure to a variety of exogenous agents, including PMA, okadaic acid, bacterial lipopolysaccharide (LPS), and combinations of cytokines and receptor-mediated priming. Cytokine synthesis induced by LPS is largely refractory to the effects of 15-deoxy-Δ^{12,14}-prostaglandin J₂ (15d-PGJ₂) and troglitazone, whereas phorbol ester and okadaic acid-induced cytokine synthesis is susceptible to their action (Fig. 2).

To establish the range of concentrations over which 15d-PGJ₂ and troglitazone are effective, we performed dose titrations of PMA- and okadaic acid-induced cytokine synthesis with monocytes prepared from several unrelated donors. The half-maximal concentration for troglitazone inhibition of PMA-induced cytokine synthesis (PMA IC₅₀) was 5.9–14.8 (mean 9.2) μ M for TNF- α , 2.9–9.3 (mean 6.0) μ M for IL-1, and 3.2–18 (mean 9.1) μ M for IL-6 (Fig. 3). Similar values for the half-maximal concentration for 15d-PGJ₂ inhibition of PMA-induced synthesis were observed: 1.6–5.2 (mean 3.8) μ M 15d-PGJ₂ for TNF α , 1.2–2.9 (mean 2.0) μ M for IL-1, and 0.8–3.1 (mean 1.6) μ M for IL-6 (Fig. 3). The close correspondence between IC₅₀ values for the different cytokines suggests a common mechanism of action. Because these values fall within the range of concentrations previously reported to be necessary for efficient induction of adipocyte differentiation *in vitro*^{7–9}, cytokine inhibition and adipogenesis probably share a regulatory circuit mediated by ligand binding to PPAR- γ .

It has been reported that several NSAIDs, including indomethacin,

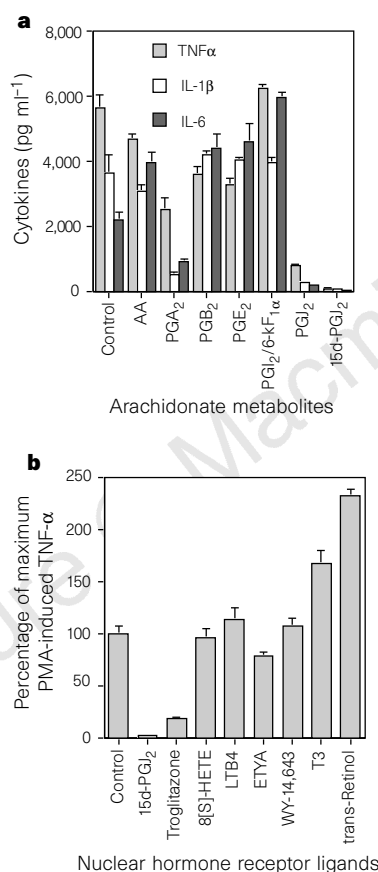


Figure 1 Effects of prostaglandin metabolites on induction by phorbol ester of inflammatory cytokines in human monocytes. **a**, Freshly prepared human monocytes were incubated in medium containing 25 nM PMA and either vehicle only (DMSO; control), 10 μ M arachidonic acid (AA), or 10 μ M of the following prostanoids: prostaglandin A₂, PGB₂, PGE₂, PGI₂/6-keto prostaglandin F_{1 α} , PGI₂, and 15d-PGJ₂. The concentrations of TNF- α , IL-1 β and IL-6 released into the medium were determined by ELISA of culture supernatants. Error bars show mean $\pm$ standard deviation of three determinations. **b**, Human monocytes were incubated in medium containing 25 nM PMA and either vehicle only (DMSO or ethanol; control), or 10 μ M 15d-PGJ₂, 30 μ M troglitazone, 10 μ M of 8[S]-hydroxy-eicosatetraenoic acid (8[S]-HETE), 3 μ M leukotriene B₄ (LTB₄), 20 μ M of eicosatetraenoic acid (ETYA), 100 μ M of WY-14643, 1 μ M of 3,3',5-triiodothyronine (T3), or 1 μ M all-*trans*-Retinol. The concentration of TNF- α released into the medium was determined by ELISA of culture supernatants. Error bars show mean $\pm$ standard deviation of three determinations normalized as a percentage of the control.

ibuprofen and fenoprofen, have PPAR- γ agonist activity at high drug concentration¹⁰. These agents are clinically important anti-rheumatic drugs, but their mechanism of action is not fully understood. As prostaglandin endoperoxide G/H synthase (cyclooxygenase) inhibitors, they relieve the pain and inflammation caused by prostaglandins synthesized within and around the inflammatory focus. However, high doses of these agents are often required to achieve significant palliation of the symptoms of rheumatoid arthritis. Such doses lead in some cases to a serum concentration greatly exceeding that necessary to inhibit cyclooxygenase types 1 or 2 (ref. 22). It has also been shown that fish-oil supplements, which are enriched for the polyunsaturated fatty acids eicosatetraenoic acid and docosahexaenoic acid, significantly reduce the symptoms of rheumatoid arthritis²³, and that docosahexaenoic acid is a naturally occurring agonist of PPAR- γ receptors²¹.

Administration of anti-TNF- γ antibodies or TNF receptor-immunoglobulin fusion proteins has been shown to ameliorate the symptoms of rheumatoid arthritis^{24–26}, and several of the NSAIDs with adipogenic activity have PPAR- γ agonist activity. It therefore seemed plausible that some of the therapeutic effect of NSAIDs might be due to the suppression of cytokine synthesis. To explore this possibility, we evaluated the ability of indomethacin, ibuprofen and fenoprofen to inhibit phorbol ester-induced production of cytokines. All three agents inhibit PMA-induced cytokine synthesis with essentially cytokine-independent dose-response characteristics (Fig. 4). The PMA IC₅₀ values for inhibition of TNF- α synthesis by indomethacin, fenoprofen and ibuprofen were 53, 265 and 485 μ M, respectively (Fig. 4, and data not shown) values that agree well with the potency and rank order of the same agents in the adipogenesis assay¹⁰. Similar experiments using okadaic acid as an inducer (Fig. 5) gave the same rank order for efficacy (15d-PGJ₂ > troglitazone > indomethacin > fenoprofen > ibuprofen) with approximate IC₅₀ values of 2, 10, 47, 133 and 142 μ M, respectively. High-dose NSAID therapy is expected to produce plasma concentrations of up to 10 μ M for indomethacin, 300 μ M for ibuprofen, and 250 μ M for fenoprofen²⁷.

Although the molecular mechanism of inhibition of cytokine synthesis by PPAR- γ agonists has yet to be fully established, the principal effects seem to be mediated pretranslationally. RNA blot analysis shows that PMA-induced accumulation of TNF- α mRNA in human peripheral blood monocytes is inhibited by PPAR- γ agonists (Fig. 6), and the approximate IC₅₀ for inhibition of TNF- α mRNA synthesis by troglitazone is 10 μ M (Fig. 6).

To determine whether changes in mRNA levels can be ascribed, at

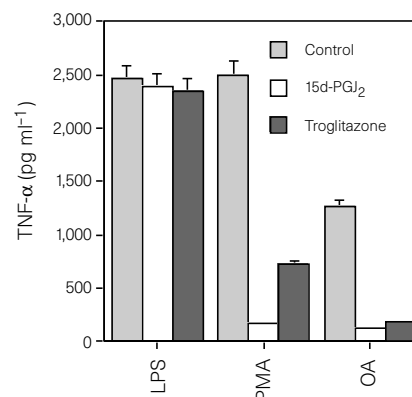


Figure 2 Effects of troglitazone and 15d-PGJ₂ on induction of TNF- α by lipopolysaccharide, phorbol ester and okadaic acid. Freshly prepared human monocytes were co-incubated with either 30 μ M troglitazone or 10 μ M 15d-PGJ₂ in medium containing 1 ng ml⁻¹ LPS, 25 nM PMA, or 50 nM okadaic acid (OA). TNF- α released into the medium was measured by ELISA. Control cultures were incubated with inducer and vehicle (DMSO) only. Error bars show mean $\pm$ standard deviation of three determinations.

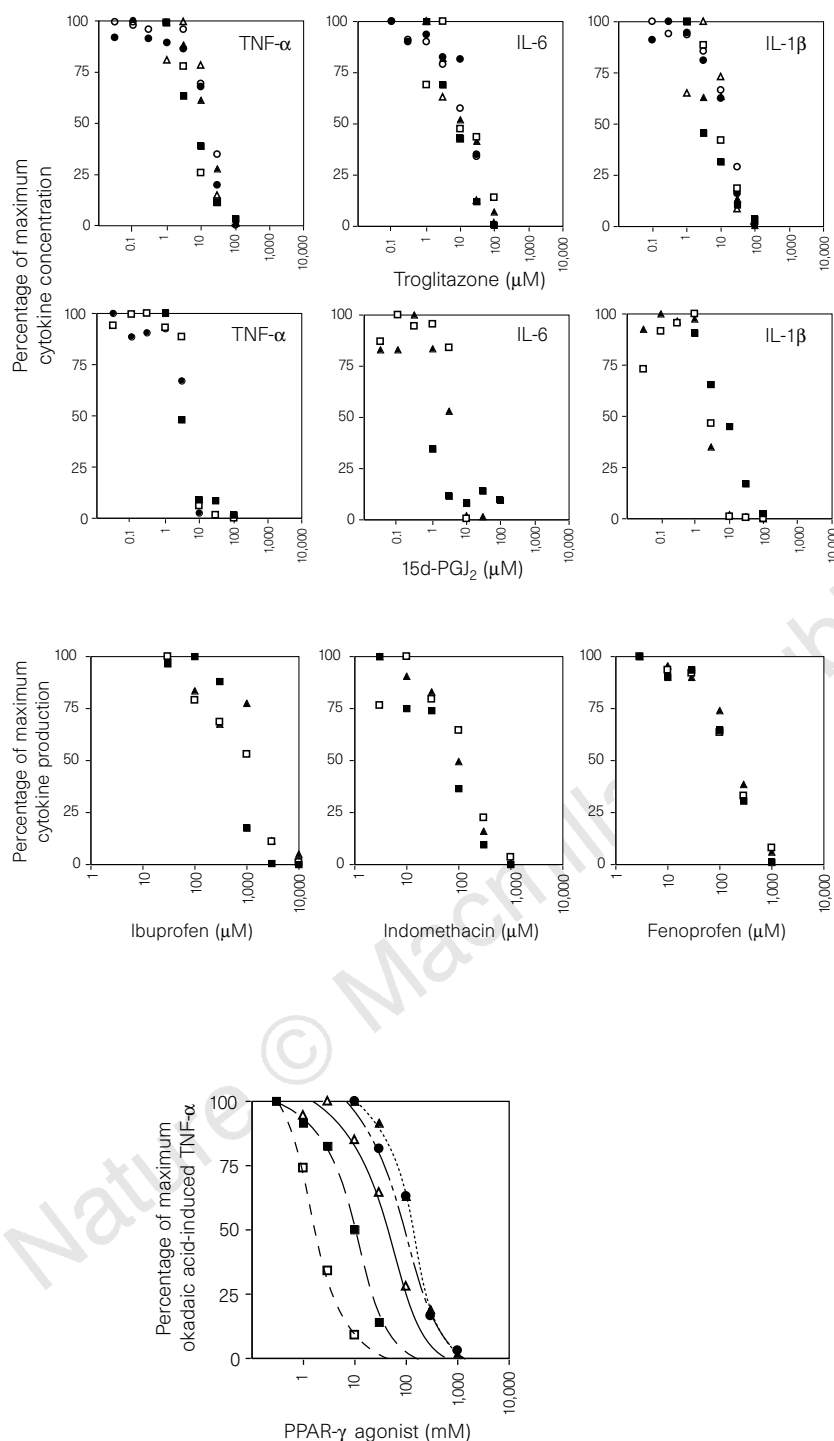


Figure 3 Dose-dependent inhibition by troglitazone and 15d-PGJ₂ of PMA-induced cytokine synthesis in monocytes prepared from several donors. Human monocytes cultures initiated from six unrelated donors were treated with 25 nM PMA and different concentrations of troglitazone (top) or 15d-PGJ₂ (bottom). The concentration of TNF- α (left), IL-6 (middle) and IL-1 β (right) released in the medium was measured by ELISA. Results are the means of three determinations, shown as the percentage of maximum cytokine concentration recorded. For clarity, standard deviations are not shown.

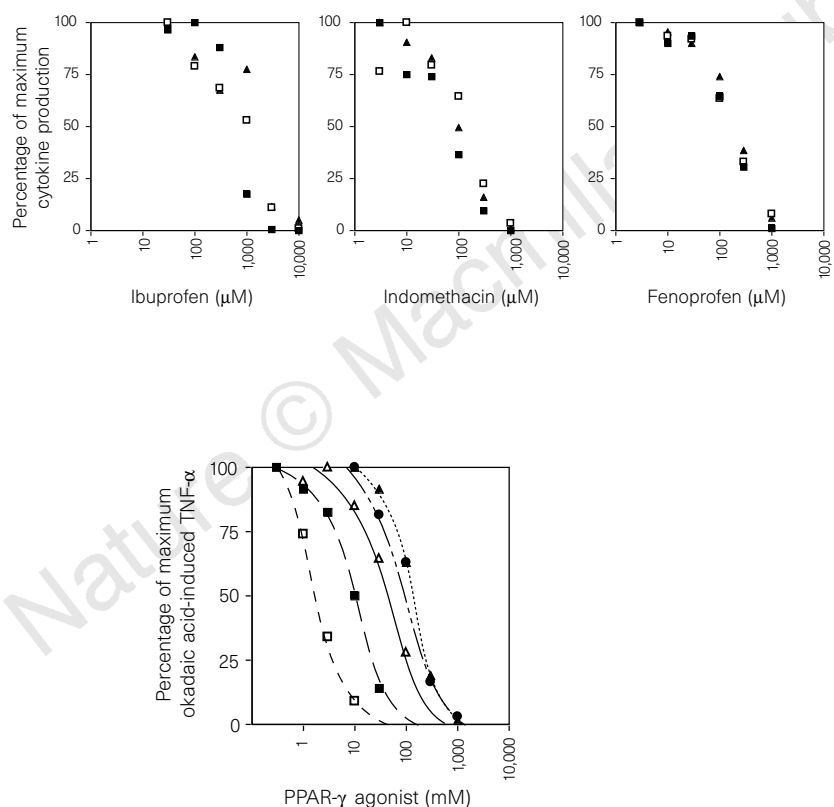


Figure 4 NSAIDs that act on PPAR- γ block production of inflammatory cytokines. Human monocyte cultures were treated with 25 nM PMA and different concentrations of ibuprofen (left), indomethacin (middle) or fenoprofen (right). Results are the means of three determinations, shown as the percentage of maximum cytokine concentration recorded.

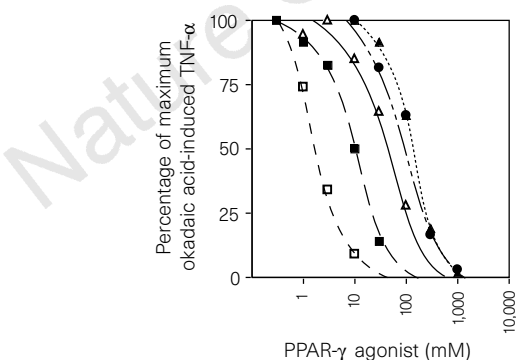


Figure 5 Dose-dependent inhibition of TNF- α production elicited by okadaic acid. Monocyte cultures were induced with okadaic acid in the presence of PPAR- γ agonists 15d-PGJ₂ (open squares), troglitazone (filled squares), indomethacin (open triangles), fenoprofen (filled circles) and ibuprofen (filled triangles) at the indicated concentrations. Results are the means of three determinations, shown as the percentage of maximum cytokine concentration recorded.

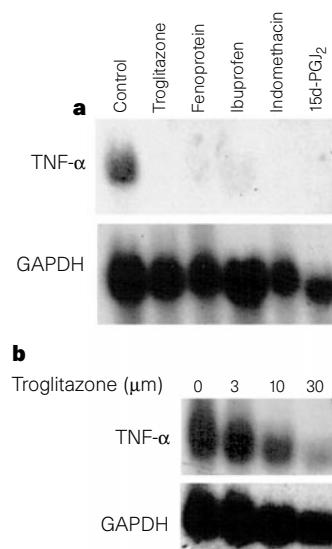


Figure 6 PMA-induced TNF- α expression in human monocytes is inhibited by PPAR- γ agonists. **a**, RNA blot analysis of human monocytes stimulated with 25 nM PMA in the presence of vehicle (DMSO; control), 100 μ M troglitazone, 1 mM fenoprofen, 1 mM ibuprofen, 1 mM indomethacin or 30 μ M 15d-PGJ₂ probed with labelled TNF- α (top) or GAPDH (bottom) cDNA. **b**, RNA blot analysis of human monocytes stimulated with 25 nM PMA in the presence of the indicated concentrations of troglitazone.

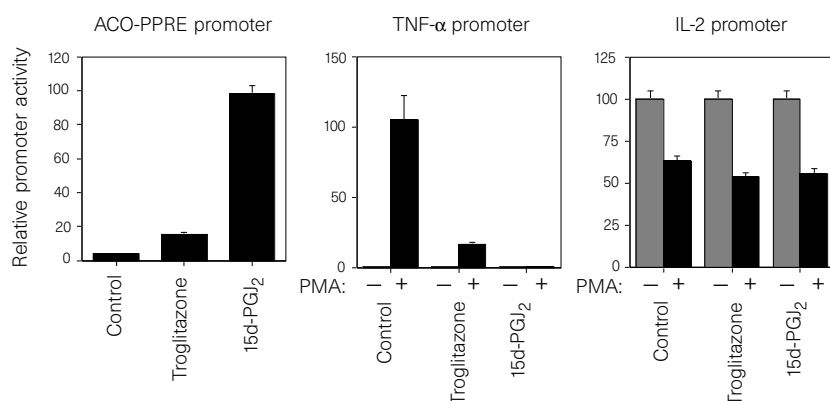


Figure 7 PPAR- γ agonists specifically inhibit TNF- α promoter-regulated gene expression. U937 cells transfected with acyl CoA oxidase-PPRE (left), TNF- α promoter (middle) and IL-2 promoter (right) luciferase reporter plasmids were stimulated with or without 25 nM PMA in the presence of either vehicle (DMSO; control), 30 μ M troglitazone or 30 μ M 15d-PGJ₂. Results are the means of three determinations normalized to the values in the non-stimulated samples.

least in part, to differences in promoter activity, we transfected the human monoblast-like cell line U937 (which, like monocytes, expresses PPAR- γ ¹⁹) with various luciferase reporter constructs. The PMA responsiveness of a truncated TNF- α promoter-luciferase reporter in transfected U937 cells is significantly inhibited by 15d-PGJ₂ and troglitazone, whereas the activity of a synthetic PPAR response element (PPRE) promoter-luciferase reporter is increased under similar conditions (Fig. 7), indicating that the effects mediated by PPAR- γ agonists vary according to the promoter. The potency of 15d-PGJ₂ in the transfection assay was substantially greater than that of troglitazone, suggesting that 15d-PGJ₂ may act through additional pathways not affected by troglitazone. The activity of the IL-2 promoter, a cytokine promoter whose output is not significantly induced by phorbol ester in these cells, was unaffected by either troglitazone or 15d-PGJ₂ (Fig. 7). Preliminary data indicate that the inhibitory action of PPAR- γ agonists in responsive cell types can also be demonstrated with synthetic promoters that have multimerized AP-1 sites (A.T., unpublished data).

Corticosteroids are potent agents for the treatment of rheumatoid arthritis, although an elicited resistance to their action ultimately limits their clinical effectiveness. Although a complete understanding of the mechanism of steroid effectiveness has yet to be elucidated, the ability of steroids to antagonize AP-1-like promoter elements in proximate cell types suggests that inflammatory stimuli acting through AP-1 and related complexes may cause rheumatoid synovitis. Understanding the pathways leading to their activation may expose new targets for the development of future disease-modifying antirheumatic drugs.

Although rheumatoid arthritis is considered to be an autoimmune disorder provoked by T-cell recognition of one or more major histocompatibility complex (MHC) class II-restricted antigens, the synovial fluid of affected individuals lacks the cytokines typically associated with T-cell proliferation²⁸; instead, cytokines associated with macrophage and fibroblast activation predominate. It is also possible that fibroblasts, which can also express PPAR- γ , are significant targets of NSAID-mediated cytokine suppression.

The existence of a PPAR- γ -dependent regulatory circuit inhibiting cytokine production is consistent with a general dichotomy between anabolism and inflammation. The inflammatory cytokine TNF- α antagonizes the synthesis of PPAR- γ , blocks adipocyte differentiation, and contributes to insulin resistance. PPAR- γ agonists in turn promote insulin sensitivity and adipocyte differentiation and block TNF- α production. Stimulation with LPS, which is presumably an acute event, appears to override the regulatory influence of PPAR- γ agonists. It therefore seems likely that the antagonism between anabolism and inflammation occurs largely in chronic processes such as parasitaemias, neoplasms and autoimmunity. □

Methods

Monocyte preparation. Human monocytes were isolated from freshly collected buffy-coat preparations of whole human blood. The mononuclear

cell fraction was prepared by dilution with an equal volume of phosphate buffered saline (PBS) at room temperature and layered over a solution containing 3 ml Ficoll-Hypaque per 10 ml blood and PBS and centrifuged for 30 min in a Sorvall RT6000 centrifuge at 900g. The mononuclear cell layer was transferred to a fresh tube, mixed with 3 volumes PBS, and centrifuged for 10 min at 400g. The supernatant was removed and the dilution and centrifugation were repeated three times. Mononuclear cells were resuspended in RPMI medium 1640 (Life Technologies, Gaithersburg, MD), counted and diluted to 5×10^6 cells per ml, after which 1 ml was transferred to each well of a 24-well tissue-culture plate and incubated for 1 h at 37 °C in a 5% CO₂ humidified incubator. The non-adherent cells were removed and the monocytes washed once with PBS before adding 1 ml fresh RPMI medium 1640 with 10% fetal bovine serum (FBS). Experiments were initiated on the day blood was collected, and all manipulations were carried out under endotoxin-free conditions.

Cytokine assay. Monocytes in fresh medium were treated with inducers and candidate induction inhibitors at the time of culture initiation. Medium was collected from triplicate wells 18–20 h after the test compounds were added. Supernatant concentrations of human TNF- α , IL-6 and IL-1 β were measured by enzyme-linked immunosorbent assay (ELISA) using protocols supplied by the manufacturers (Immunotech and Endogen, Boston, MA). To obtain IC₅₀ values, data were fit to a four-parameter exponential and the derived parameters used to calculate the concentration at which 50% of the maximal activity was observed.

RNA blot. Total RNA of human monocyte cultures was isolated 15 h after stimulation with 25 nM PMA in the absence or presence of PPAR- γ agonists. RNA blot analyses were performed with standard procedures and labelled probes prepared from human TNF- α and GAPDH cDNA with the Megaprime DNA labelling kit (Amersham Life Science) according to the manufacturer's instructions.

Luciferase assay. U937 cells (2×10^7) were transfected with 10–20 μ g luciferase reporter DNA by electroporation at 875 V cm⁻¹, 960 μ F (Bio-Rad Laboratories). Transfected cells were allowed to recover for 1 h and triplicate samples were either untreated or treated with 25 nM PMA in the absence or presence of the indicated drugs. Luciferase activity was assayed 18–36 h later using the Dual-luciferase reporter assay system (Promega) with the pRL-TK vector as the internal reporter control. TNF- α luciferase reporter contains the human TNF- α promoter from positions -283 to +113. The ACO-PPRE luciferase reporter contains three copies of PPRE from the promoter of rat acyl CoA oxidase. The IL-2 luciferase reporter contains the human IL-2 promoter from positions -577 to +53.

Received 5 September; accepted 14 October 1997.

1. Mangelsdorf, D. J. *et al.* The nuclear receptor superfamily: the second decade. *Cell* **83**, 835–839 (1995).
2. Lemberger, D., Desvergne, B. & Wahli, W. Peroxisome proliferator-activated receptors: a nuclear receptor signaling pathway in lipid physiology. *Annu. Rev. Cell Dev. Biol.* **12**, 335–363 (1996).
3. Chawla, A., Schwarz, E. J., Dimaculangan, D. D. & Lazar, M. A. Peroxisome proliferator-activated receptor (PPAR) gamma: adipose-predominant expression and induction early in adipocyte differentiation. *Endocrinology* **135**, 798–800 (1994).
4. Tontonoz, P., Hu, E., Graves, R. A., Budavari, A. I. & Spiegelman, B. M. mPPAR gamma 2: tissue-specific regulator of an adipocyte enhancer. *Genes Dev.* **8**, 1224–1234 (1994).
5. Tontonoz, P., Hu, E. & Spiegelman, B. M. Stimulation of adipogenesis in fibroblasts by PPAR gamma 2, a lipid-activated transcription factor. *Cell* **79**, 1147–1156 (1994).
6. Hu, E., Tontonoz, P. & Spiegelman, B. M. Transdifferentiation by myoblasts by the adipogenic transcription factor PPAR gamma and C/EBP alpha. *Proc. Natl Acad. Sci. USA* **92**, 9856–9860 (1995).
7. Lehmann, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome

- proliferator-activated receptor gamma (PPAR gamma). *J. Biol. Chem.* **270**, 12953–12956 (1995).
8. Kliewer, S. A. *et al.* A prostaglandin J2 metabolite binds peroxisome proliferator-activated receptor gamma and promotes adipocyte differentiation. *Cell* **83**, 813–819 (1995).
 9. Forman, B. M. *et al.* 15-Deoxy-delta 12,14-prostaglandin J2 is a ligand for the adipocyte determination factor PPAR gamma. *Cell* **83**, 803–812 (1995).
 10. Lehmann, J. M., Lenhard, J. M., Oliver, B. B., Ringold, G. M. & Kliewer, S. A. Peroxisome proliferator-activated receptors alpha and gamma are activated by indomethacin and other non-steroidal anti-inflammatory drugs. *J. Biol. Chem.* **272**, 3406–3410 (1997).
 11. Hotamisligil, G. S., Shargill, N. S. & Spiegelman, B. M. Adipose expression of tumor necrosis factor-alpha: direct role in obesity-linked insulin resistance. *Science* **259**, 87–91 (1993).
 12. Patton, J. S. *et al.* Interferons and tumor necrosis factors have similar catabolic effects on 3T3 L1 cells. *Proc. Natl Acad. Sci. USA* **83**, 8313–8317 (1986).
 13. Torti, F. M., Torti, S. V., Larrick, J. W. & Ringold, G. M. Modulation of adipocyte differentiation by tumor necrosis factor and transforming growth factor beta. *J. Cell Biol.* **108**, 1105–1113 (1989).
 14. Marshall, M. K., Doerrler, W., Feingold, K. R. & Grunfeld, C. Leukemia inhibitory factor induces changes in lipid metabolism in cultured adipocytes. *Endocrinology* **135**, 141–147 (1994).
 15. Berg, M., Fraker, D. L. & Alexander, H. R. Characterization of differentiation factor/leukaemia inhibitory factor effect on lipoprotein lipase activity and mRNA in 3T3-L1 adipocytes. *Cytokine* **6**, 425–432 (1994).
 16. Lang, C. H., Dobrescu, C. & Bagby, G. J. Tumor necrosis factor impairs insulin action on peripheral glucose disposal and hepatic glucose output. *Endocrinology* **130**, 43–52 (1992).
 17. Szalkowski, D., White-Carrington, S., Berger, J. & Zhang, B. Antidiabetic thiazolidinediones block the inhibitory effect of tumor necrosis factor-alpha on differentiation, insulin-stimulated glucose uptake, and gene expression in 3T3-L1 cells. *Endocrinology* **136**, 1474–1481 (1995).
 18. Braissant, O., Foulle, F., Scotto, C., Dauca, M. & Wahli, W. Differential expression of peroxisome proliferator-activated receptors (PPARs): tissue distribution of PPAR-alpha, -beta, and -gamma in the adult rat. *Endocrinology* **137**, 354–366 (1996).
 19. Greene, M. E. *et al.* Isolation of the human peroxisome proliferator activated receptor gamma cDNA: expression in hematopoietic cells and chromosomal mapping. *Gene Express.* **4**, 281–299 (1995).
 20. Devchand, P. R. *et al.* The PPARalpha-leukotriene B4 pathway to inflammation control. *Nature* **384**, 39–43 (1996).
 21. Yu, K. *et al.* Differential activation of peroxisome proliferator-activated receptors by eicosanoids. *J. Biol. Chem.* **270**, 23975–23983 (1995).
 22. Meade, E. A., Smith, W. L. & DeWitt, D. L. Differential inhibition of prostaglandin endoperoxide synthase (cyclooxygenase) isozymes by aspirin and other non-steroidal anti-inflammatory drugs. *J. Biol. Chem.* **268**, 6610–6614 (1993).
 23. Lau, C. S., Morley, K. D. & Belch, J. J. Effects of fish oil supplementation on non-steroidal anti-inflammatory drug requirement in patients with mild rheumatoid arthritis—a double-blind placebo controlled study. *Br. J. Rheumatol.* **32**, 982–989 (1993).
 24. Feldmann, M., Brennan, F. M. & Maini, R. N. Role of cytokines in rheumatoid arthritis. *Annu. Rev. Immunol.* **14**, 397–440 (1996).
 25. Brennan, F. M. & Feldman, M. Cytokines in autoimmunity. *Curr. Opin. Immunol.* **8**, 872–877 (1996).
 26. Moreland, L. W. Treatment of rheumatoid arthritis with a recombinant human tumor necrosis factor receptor (p75)-Fc fusion protein. *N. Engl. J. Med.* **337**, 141–147 (1997).
 27. McEvoy, G. K. (ed) *AHFS 97 Drug Information* (American Society of Health-Systems Pharmacists, Bethesda, MD, 1997).
 28. Firestein, G. S. & Zvaifler, N. J. Anticytokine therapy in rheumatoid arthritis. *N. Engl. J. Med.* **337**, 195–197 (1997).

Acknowledgements. We thank T. Gulick for advice, guidance, time and key reagents. This work was supported by a grant from Amgen, Inc., and awards from the NIH. A.T. is a fellow of the Cancer Research Institute.

Correspondence and requests for materials should be addressed to B.S. (e-mail: seed@molbio.mgh.harvard.edu).

G-protein-coupled receptor of Kaposi's sarcoma-associated herpesvirus is a viral oncogene and angiogenesis activator

Carlos Bais*†, Bianca Santomasso*†, Omar Coso‡, Leandros Arvanitakis§, Elizabeth Geras Raaka‡, J. Silvio Gutkind‡, Adam S. Asch†, Ethel Cesarman§, Marvin C. Gerhengorn‡ & Enrique A. Mesri*†

* Laboratory of Viral Oncogenesis, † Division of Hematology–Oncology, and ‡ Division of Molecular Medicine, Department of Medicine, and § Department of Pathology, Cornell University Medical College, New York, New York 10021, USA
‡ Molecular Signaling Unit, Laboratory of Cellular Development and Oncology, NIDR, NIH, Bethesda, Maryland 20892, USA

The Kaposi's sarcoma-associated herpesvirus (KSHV/HHV8) is a γ -2 herpesvirus^{1–5} that is implicated in the pathogenesis of Kaposi's sarcoma^{1,5} and of primary effusion B-cell lymphomas (PELs)⁶. KSHV infects malignant and progenitor cells of Kaposi's sarcoma⁷ and PEL^{2,6,8}, it encodes putative oncogenes^{4,5,9} and genes that may cause Kaposi's sarcoma pathogenesis by stimulating

angiogenesis^{4,5,9,10}. The G-protein-coupled receptor encoded by an open reading frame (ORF 74) of KSHV⁹ is expressed in Kaposi's sarcoma lesions and in PEL^{9,11} and stimulates signalling pathways linked to cell proliferation¹² in a constitutive (agonist-independent) way¹². Here we show that signalling by this KSHV G-protein-coupled receptor leads to cell transformation and tumorigenicity, and induces a switch to an angiogenic phenotype¹³ mediated by vascular endothelial growth factor¹⁴, an angiogenesis^{13,14} and Kaposi's-spindle-cell growth factor^{15–17}. We find that this receptor can activate two protein kinases, JNK/SAPK and p38MAPK, by triggering signalling cascades like those induced by inflammatory cytokines¹⁸ that are angiogenesis activators¹⁹ and mitogens for Kaposi's sarcoma cells¹⁰ and B cells. We conclude that the KSHV G-protein-coupled receptor is a viral oncogene that can exploit cell signalling pathways to induce transformation and angiogenesis in KSHV-mediated oncogenesis.

Persistently activated G-protein-coupled receptors (GPCRs) can transform cells²⁰ and act as oncogenes in human cancers²¹. To determine whether the constitutively active KSHV-GPCR was oncogenic, we used a focus-formation assay in NIH3T3 cells. Transfection of NIH3T3 cells with the expression vector pCEFL-KSHV-GPCR (pCKSHV-GPCR) led to dose-dependent appearance of foci of transformation three weeks after transfection, whereas cells transfected with pCEFL vector did not form any foci (Table 1). The pCEFL and pCKSHV-GPCR constructs showed similar transfection efficiency but only pCKSHV-GPCR induced foci of transformation; the transforming potency of KSHV-GPCR was comparable to that of the m1 muscarinic receptor (m1 AChR), and agonist-dependent transforming GPCR²⁰ (Table 1). All pCKSHV-GPCR-transformed cells and all pCKSHV-GPCR transiently or stably transfected cells expressed KSHV-GPCR, as determined by polymerase chain reaction with reverse transcription (RT-PCR) (not shown), and accumulated significantly more inositol phosphate (InsP) second messenger molecules¹² than did pCEFL-transfected cells. Six NIH3T3 cell clones stably expressing KSHV-GPCR constitutively produced 2.8 ± 1.6 -fold more InsP than did pCEFL-transfected clones ($P < 0.05$), and two neo-enriched populations of cells expressing KSHV-GPCR produced InsP at rates that were sixfold faster than pCEFL-transfected populations. These results indicate that in all these transfected NIH3T3 cells, the KSHV-GPCR gene was not only expressed but its gene product was biologically active and signalling. To show that cell transformation is a result of KSHV-GPCR signalling, we investigated the effects of GPCR-specific kinases (GRKs), which phosphorylate and inhibit activated GPCRs²², on their signalling. We found that GRK-5, a GRK that can desensitize KSHV-GPCR signalling²³, gave up to a 60% inhibition of focus formation (Fig. 1), whereas GRK-2, which

Table 1 Cell-transforming and tumorigenic activity of KSHV-GPCR

Plasmid	Neo ^r colonies*	Foci of cell transformation†	Tumour forming clones‡
pCEFL	3,500	0	0/3
pCEFL-KSHV-GPCR	2,800	400	3/4
pCMV-m1 AChR§	3,000	680	ND
pCEFL-RasV12	2,500	2,000	ND

NIH3T3 cells was transfected with 0.2 or 2 μ g of plasmid DNA as described in Methods. ND, not determined.

* Efficiency of transfection was determined by plating the transfected cells in medium containing 750 μ g ml⁻¹ of G418.

† Foci of transformation were scored after 3 weeks for the GPCRs and after 10 d for RasV12. Formation of foci was dependent on the dose of transfected DNA; focus formation potency was determined by extrapolating the data from 0.2 μ g transfected DNA. Results represent the median of three experiments. The morphology of foci obtained with KSHV-GPCR was similar to m1 AChR-induced foci (see Supplementary Information).

‡ Transformed cells from representative foci selected in neomycin, or cells stably transfected with pCEFL, were injected into BALB/c nude mice ($n = 4$). Tumour formation was scored after a month. The three tumorigenic clones produced tumours in all four mice injected. The three pCEFL stably transfected clones did not produce tumours in any of the four mice injected.

§ The m1 AChR (pCMVm1) was used as a positive control for GPCR-induced transformation²⁰ after transfection cells were cultured in the presence of 100 μ M carbachol²⁰.

|| RasV12 (pCEFL RasV12) was used as a positive control for agonist-independent focus formation.

|| Present addresses: Molecular Virology Laboratory, Hellenic Pasteur Institute, Athens GR-115 21, Greece (L.A.); Laboratorio de Fisiología y Biología Molecular, Dept de Cs. Biológicas, F.C.E. y N. Universidad de Buenos Aires, Buenos Aires, Argentina (O.C.).

does not affect KSHV-GPCR signalling, did not affect focus formation even at high doses (Fig. 1). This indicates that transformation is a consequence of signalling triggered by KSHV-GPCRs.

Focus-derived transformed cells were selected in neomycin and injected into the flank of nude mice. Three of the four tested KSHV-GPCR-transformed clones caused tumours (Table 1, Fig. 2a); the time taken for a 1.5-cm³ tumour to form ranged from 10 days to 4 weeks, whereas none of the pCEFL stable transfectants were tumorigenic (Table 1, Fig. 2b). The tumours had the histomorphological features of fibrosarcomas (Fig. 2c) and were formed by plump spindle cells with relatively abundant eosinophilic cytoplasm. Cells were atypical, with vesicular nuclei and prominent nucleoli, and mitotic figures were easily identified. Cells removed from freshly dissected tumours and selected in neomycin expressed KSHV-GPCR and produced large amounts of inositol phosphate (not shown), indicating that a strong KSHV-GPCR activity was associated with the tumorigenic state. These results show that in NIH3T3 cells, KSHV-GPCR is an oncogene that triggers intracellular signalling cascades leading to cell transformation and tumorigenicity.

KSHV-GPCR is most homologous to the human receptor for the angiogenic chemokine interleukin-8 (ref. 24), and we showed

previously that KSHV-GPCR signalling stimulates transcription controlled by AP-1-containing promoters¹², such as those controlling the expression of several angiogenic growth factors²⁵. KSHV-GPCR could, therefore, induce angiogenesis or evoke angiogenic-like signalling. KSHV-GPCR-transformed cells formed vascularized tumours (Fig. 2b), indicating that either the cells were angiogenic before the mouse injection, or that they went through an angiogenic switch¹³ while growing *'in vivo'*. As angiogenicity cannot be correlated with KSHV-GPCR expression in transformed cells because they can undergo further genetic changes that affect angiogenicity, we tested whether KSHV-GPCR induces angiogenicity using transiently or stably transfected NIH3T3 cells. We tested conditioned medium from pCKSHV-GPCR-transfected cells and controls for ability to induce endothelial cell growth and angiogenesis *'in vitro'*. Figure 3a, b shows that in all states post-transfection, and for all stably transfected clones, media conditioned by KSHV-GPCR-expressing cells could stimulate growth of human umbilical vein endothelial cells (HUVECs). In contrast, conditioned medium from cells and clones transfected with pCEFL, like untransfected NIH3T3 cells, could not support endothelial cell growth (Fig. 3a, b). To confirm the angiogenicity *in vitro* of pCKSHV-GPCR-transfected

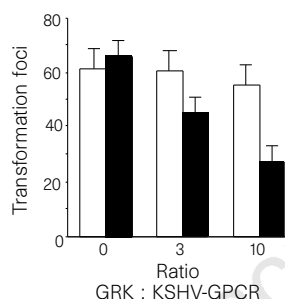


Figure 1 Inhibition of KSHV-GPCR-induced focus formation by GRK-5. Focus-forming activity of KSHV-GPCR in NIH3T3 cells in the presence of GRK-5 and GRK-2. NIH3T3 cells were transfected with 0.2 µg pCKSHV-GPCR alone or with 0.6 µg or 2 µg of GRK-5 (black bars) or GRK-2 (white bars) plasmid. Data represent the mean ± range of duplicate determinations in one representative experiment.

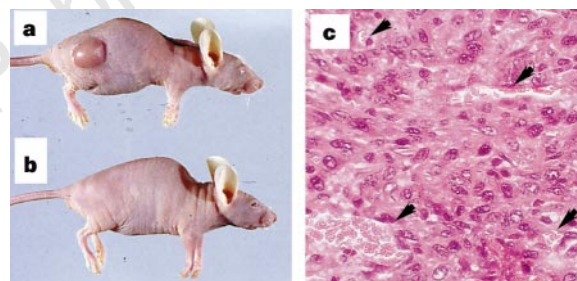
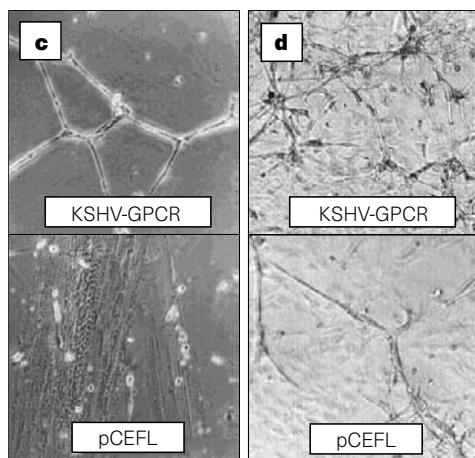
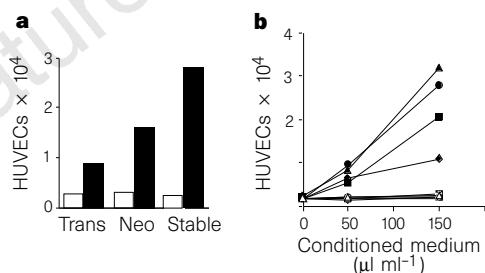


Figure 2 Tumorigenicity of KSHV-GPCR-transformed NIH3T3 cells in nude mice. **a, b**, Photographs of mice taken 10 d after injection with cells **a**, from KSHV-GPCR-induced focus of transformation, or **b**, stably transfected with pCEFL vector. **c**, Microscopic appearance of a tumour stained with haematoxylin-eosin (original magnification, ×400). Arrows indicate blood vessels.

Figure 3 Angiogenic response *in vitro* induced by NIH3T3 cells transfected with KSHV-GPCR. **a, b**, Paracrine stimulation of endothelial cell growth (HUVECs). **a**, Mitogenicity to HUVECs of conditioned medium from NIH3T3 cells transfected with pCKSHV-GPCR (black bars) or pCEFL (white bars). Trans, transient transfection; Neo, cell populations enriched in transfectants by mass selection in neomycin; Stable, stable transfectants. **b**, Mitogenicity to HUVECs of conditioned medium from KSHV-GPCR (filled symbols) or pCEFL (open symbols) stably transfected clones and non-transfected NIH3T3 cells (open triangles). Experiments show the mean of duplicate determinations in one representative experiment; all duplicates deviated from their means by less than 10%. **c**, Microtubule formation *in vitro* induced by conditioned medium from transfected cells. Photographs show the appearance of the matrigel surface 24 h after seeding the HUVECs in the presence of different conditioned media (phase contrast; original magnification, ×100). Note the formation of interconnected microtubules in the endothelial cells incubated with KSHV-GPCR-conditioned medium. **d**, Microtubule formation *in vitro* in matrigel by cocultivation of HUVECs with transfected cells. Transfected cells were incubated with HUVECs seeded in the top of the matrigel; note the increase in microtubule formation and the invasion of the tridimensional matrix by HUVECs incubated with KSHV-GPCR-transfected cells (bottom of the plate/background cells). Photographs show the appearance of the matrigel surface 48 h after incubation (phase contrast; original magnification, ×100).

clones, we tested their conditioned media in a microtubule-formation assay²⁶ and in a co-culture matrigel assay²⁷. Figure 3c shows that conditioned medium from pCKSHV-GPCR-transfected cells induced formation of endothelial cell microtubules in matrigel, whereas control medium did not. HUVECs seeded over a matrigel layer covering the cells transfected with pCKSHV-GPCR (Fig. 3d) invaded the matrigel to form a network of interconnecting microtubules²⁷, whereas pCEFL-transfected cells induced background microtubule formation on top of the matrigel layer (Fig. 3d). These results indicate that expression of KSHV-GPCR is sufficient to produce an imbalance between secreted angiogenic inducers and inhibitors which leads to a switch to an angiogenic phenotype¹³. Vascular endothelial growth factor is a major angiogenic inducer^{13,14} and KS-spindle-cell growth factor^{15–17} that can be transcriptionally regulated by the AP-1-responsive promoter in the VEGF gene²⁵. Figure 4a shows that conditioned medium from cells transfected with pCKSHV-GPCR contained from two (transient transfection) to six (neomycin-enriched and stable transfectants) times more VEGF than did medium from pCEFL-transfected cells (Fig. 4a) or non-transfected NIH3T3 cells (not shown). VEGF levels strongly correlated with the stimulation of endothelial cell growth ($r = 0.92$), suggesting that VEGF might mediate this mitogenic response. To test whether angiogenic responses *in vitro* are mediated by VEGF, we used an anti-mouse-VEGF antibody, which specifically blocks VEGF activity. As shown in Fig. 4d, anti-VEGF antibody inhibited responses induced by

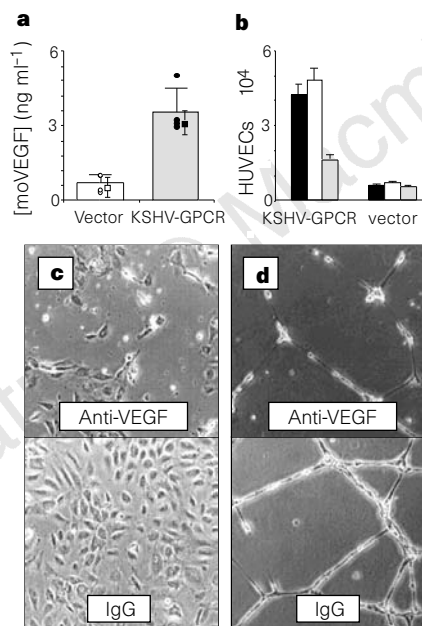


Figure 4 The angiogenic response induced by KSHV-GPCR is mediated by VEGF. **a**, Levels of VEGF measured by ELISA. Results are the mean of duplicate determinations in one representative experiment; duplicates deviated from their means by less than 5%. Cells transfected with pCEFL are represented by white circles; cells transfected with KSHV-GPCR by black circles. **b**, **c**, Anti-VEGF antibody inhibits endothelial cell mitogenicity of media conditioned by KSHV-GPCR-transfected NIH3T3 cells. **b**, Conditioned media were incubated with no addition (black bars), preimmune goat total IgG (white bars) or anti-mouse VEGF antibody (grey bars). Data represent the mean $\pm$ range of duplicate determinations in one representative experiment. **c**, Appearance of endothelial cells after incubation with conditioned medium two days after plating (phase contrast; original magnification, $\times 100$). **d**, Anti-VEGF antibody inhibits *in vitro* microtubule-formation activity of media conditioned by KSHV-GPCR-transfected NIH3T3 cells. Before assay, the conditioned media were incubated with the antibodies indicated. Photographs show the appearance of the microtubules after 24 h incubation (phase contrast; original magnification, $\times 100$).

conditioned medium from pCKSHV-GPCR-transfected cells both in the endothelial cell growth assay and in the microtubule formation assay (Fig. 4b–d). These results show that VEGF, a well characterized mediator of physiological and pathological angiogenesis, and a KS-spindle-cell growth factor, also controls the KSHV-GPCR-induced angiogenic switch.

Our results indicate that this KSHV gene subverts cell signalling pathways leading to oncogenesis and angiogenicity. The three principal kinase cascades that mediate signalling between membrane receptors and the nucleus in mammalian cells are the ERK-2/MAPK kinase cascade induced by mitogen activation of tyrosine-kinase receptors and Ras, and the JNK/SAPK and p38MAPK pathways induced by inflammatory cytokines and stress¹⁸. To investigate the signalling pathways activated by KSHV-GPCR, we expressed epitope-tagged MAPK/ERK-2, JNK/SAPK or p38MAPK together with KSHV-GPCR in HEK 293T cells and found that KSHV-GPCR could activate JNK/SAPK and p38MAPK but not ERK-2/MAPK (Fig. 5); these data indicate that constitutive signalling by KSHV-GPCR can recruit protein kinase pathways characteristic of activation by inflammatory cytokines¹⁸. These results are consistent with the idea that KSHV-GPCR signalling in KSHV-infected cells can mimic signalling by agonists that cause proliferation of Kaposi's sarcoma and B cells, and that are Kaposi's-spindle-cell angiogenesis activators¹⁹.

We have presented three lines of evidence indicating that KSHV-GPCR can participate in KSHV-mediated oncogenesis. We found that KSHV-GPCR, which is expressed in KSHV-associated malignancies^{9,11}, can transform cells and induce tumorigenicity. It is still not known whether KSHV-mediated transformation occurs as with Epstein-Barr virus (EBV/HHV-4), as a consequence of genetic modifications in immortalized latently infected proliferating cells²⁸, or as an event occurring during productive infection as a result of an abortive lytic cycle. KSHV-GPCR is an early lytic gene

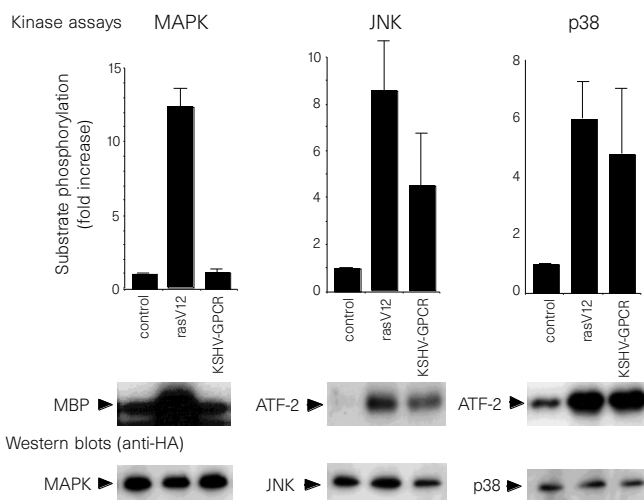


Figure 5 KSHV-GPCR can activate members of the MAPK superfamily. HEK 293T cells were transfected with pcDNA3-HA-ERK2, pcDNA3-HA-JNK or pCEFL-HA-p38 (1 μ g per plate), together with pcDNA3 vector (control) or with pCKSHV-GPCR as indicated (2 μ g per plate in each case). As positive controls, cells were co-transfected with an activated mutant of *ras* (pCEFL RasV12) or stimulated with anisomycin (10 μ g ml⁻¹) for ERK2 (MAPK) or JNK and p38, respectively. Kinase reactions were performed on anti-HA immunoprecipitates from the corresponding lysates. A fraction of the total lysates was used for western blotting analysis. Autoradiograms correspond to representative experiments. ³²P-labelled products, as well as specific bands detected by the anti-HA antibody, are indicated by arrows. Data represent the mean $\pm$ s.e.m. of three independent experiments and are expressed as fold increase with respect to vector-transfected cells (control).

that shows basal expression in latently infected populations and is expressed in *de novo* infections (R. Nador and O. Flore, personal communications), indicating that it could participate in both types of oncogenesis. We have shown that KSHV-GPCR expression is sufficient to induce a switch to an angiogenic phenotype by secretion of VEGF, an angiogenesis growth factor^{13,14} and Kaposi's sarcoma cell mitogen. KSHV-GPCR is expressed in Kaposi's sarcoma lesions and induction of the angiogenic phenotype and VEGF secretion by KSHV-GPCR can therefore participate in autocrine or paracrine stimulation of the KS angiogenic and proliferative response. Finally, we have shown that KSHV-GPCR can activate JNK/SAPK and p38 MAPK kinases, indicating that this viral protein has the molecular specificity to trigger both signalling cascades activated by inflammatory cytokines¹⁸ that induce angiogenicity and upregulate VEGF in Kaposi's spindle cells^{17,19} and induce a Kaposi's sarcoma phenotype in endothelial cells. Our results suggest that 'signalling mimicry' by KSHV-GPCR in KSHV-infected cells could activate an inflammatory-cytokine-like signalling pathway leading to angiogenesis and point to possible mechanisms for chemokine-induced angiogenesis²⁴. To our knowledge, this is the first demonstration that a KSHV gene is capable of inducing both transformation and an angiogenic phenotype in mammalian cells, strongly supporting the concept that KSHV infection plays a direct role in Kaposi's sarcoma pathogenesis and lymphomagenesis. □

Methods

Cell transfection and focus formation assays. To express KSHV-GPCR in NIH3T3 cells (Swiss mouse), a KSHV-GPCR DNA fragment obtained from pKSHVGPCR¹² (pDNA3-KSHVGPCR, CMV promoter) was subcloned into the pCEFL plasmid (mouse EF 12 promoter), which also codes for neomycin resistance, to obtain pCEFL-KSHV-GPCR (pCKSHV-GPCR). All transfections were carried out by the calcium phosphate method²⁰. Focus formation was assayed as described²⁰. For neomycin enrichment, cells were incubated in DMEM with 10% calf serum (Bio-Whittaker) plus 750 µg ml⁻¹ of G418 (Geneticin, Gibco-BRL). To obtain clonal stable transfectants, the selected neomycin plates were split in order to obtain isolated colonies. pCKSHV-GPCR and pCEFL stable transfectants were maintained in 750 µg ml⁻¹ G418 and always passaged before confluence.

Tumorigenicity studies. Transformed cells from foci were isolated by trypsinization within a cloning cylinder, and grown in DMEM with 10% calf serum, 750 µg ml⁻¹ G418. For injection, cells were trypsinized, resuspended in PBS and injected in the right flank of 8–10-week-old female BALB/c nu/nu mice (3 × 10⁵ cells per 200 µl PBS per animal).

Inositol phosphate accumulation. InsP accumulation was determined as described¹². InsP accumulation during 90 min was calculated as the difference between InsP levels in cells incubated with or without LiCl. InsP levels were normalized as a percentage of the total ³H radioactivity in inositol phosphates plus lipids.

Endothelial cell growth assay. Fresh HUVECs (passage 4 to 6) were grown in gelatin-coated plates with M199 medium 20% FBS, 2 ng ml⁻¹ bFGF (GIBCO-BRL). The HUVECs were seeded in gelatin-free 24-well plates (10⁴ cells per plate) and the mitogenicity assay was carried out as described²⁹. Conditioned medium was obtained from confluent cultures by incubation with neomycin-free medium (with or without calf serum) during 48 h.

Microtubule formation in matrigel *in vitro*. Wells of a 24-multiwell plate were coated with 100 µl per well of liquid extracellular matrix extract (Sigma) or Matrigel (Collaborative Biomed) and incubated for 30 min at 37 °C. HUVEC cells (10⁴) in medium with 10% calf serum were added per well and conditioned medium was added once cells were attached. Plates were observed after 24 h. For co-culture experiment, matrigel was added on top of transfected cell and microtubule formation was allowed to proceed for 2–3 d.

VEGF ELISA. VEGF in the conditioned medium was measured using a mouse-VEGF Quantikine kit (R&D), following the manufacturer's recommended procedures. For calibration, we used mVEGF standards provided with the kit.

Anti-VEGF antibody inhibition of angiogenic activity *in vitro*. Conditioned media were incubated with 0.2 µg ml⁻¹ of anti-mouse-VEGF polyclonal antibody (total goat IgG; R&D) or with preimmune total goat IgG (R&D)

for 1 h room temperature. The preincubated conditioned media were used in HUVEC assays and matrigel microtubule formation assays.

Kinase assays. Phosphorylating activities of transfected, epitope-tagged MAPK, JNK and p38 MAPK were determined as described³⁰. Cleared cell lysates were immunoprecipitated with anti-HA monoclonal antibody and an *in vitro* kinase assay was performed using myelin basic protein (MBP) or GST-ATF2 as substrates for MAPK or JNK and p38, respectively.

Western blots. Western blot analysis using anti-HA epitope antibody was done using a fraction of the total lysates as described³⁰.

Received 11 July; accepted 7 October 1997.

- Chang, Y. *et al.* Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266**, 1865–1869 (1994).
- Mesri, E. A. *et al.* Human herpesvirus-8/Kaposi's sarcoma-associated herpesvirus is a new transmissible virus that infects B-cells. *J. Exp. Med.* **183**, 2385–2390 (1996).
- Renne, R. *et al.* Lytic growth of Kaposi's sarcoma associated herpesvirus (HHV-8) in culture. *Nature Med.* **2**, 342–346 (1996).
- Russo, J. L. *et al.* Nucleotide sequence of the Kaposi's sarcoma associated herpesvirus (HHV8). *Proc. Natl Acad. Sci. USA* **93**, 14862–14867 (1997).
- Boshoff, C. & Weiss, R. A. Kaposi's sarcoma-associated herpesvirus. *Curr. Opin. Infect. Dis.* **10**, 26–31 (1997).
- Cesarman, E., Chang, Y., Moore, P. S., Said, J. W. & Knowles, D. M. Kaposi's sarcoma-associated herpesvirus-like DNA sequences are present in AIDS-related body cavity B-cell lymphomas. *N. Engl. J. Med.* **332**, 1186–1191 (1995).
- Boshoff, C. *et al.* Kaposi's sarcoma-associated herpesvirus infects endothelial and spindle cells. *Nature Med.* **1**, 1274–1278 (1995).
- Arvanitakis, L. *et al.* Establishment and characterization of a body cavity-based lymphoma cell line (BC-3) harboring Kaposi's Sarcoma-associated herpesvirus (KSHV/HHV-8) in the absence of EBV. *Blood* **86**, 2708–2714 (1996).
- Cesarman, E. *et al.* Kaposi's sarcoma-associated herpesvirus contains G protein-coupled receptor and cyclin D homologs which are expressed in Kaposi's sarcoma and malignant lymphoma. *J. Virol.* **70**, 8218–8223 (1996).
- Ensolli, B., Barillari, G. & Gallo, R. C. Cytokines and growth factors in the pathogenesis of AIDS-associated Kaposi's Sarcoma. *Immunol. Rev.* **127**, 147–155 (1992).
- Guo, H. G. *et al.* Characterization of a chemokine receptor-related gene in human herpesvirus 8 and its expression in Kaposi's sarcoma. *Virology* **228**, 371–378 (1997).
- Arvanitakis, L., Geras Raaka, E., Varma, A., Gershengorn, M. & Cesarman, E. Human herpesvirus KSHV encodes a constitutively active G-protein coupled receptor linked to cell proliferation. *Nature* **385**, 347–350 (1997).
- Hanahan, D. & Folkman, J. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell* **86**, 356–364 (1996).
- Ferrara, N. Vascular endothelial growth factor. *Eur. J. Cancer* **14**, 2413–2422 (1996).
- Masood, R. *et al.* Vascular endothelial growth factor/vascular permeability factor is an autocrine growth factor in AIDS-Kaposi's sarcoma. *Proc. Natl Acad. Sci. USA* **94**, 979–984 (1997).
- Nakamura, S., Murakamimori, K., Rao, N., Weich, H. A. & Rajeev, B. Vascular endothelial growth factor is a potent angiogenic factor in aids-associated kaposi-sarcoma-derived spindle cells. *J. Immunol.* **158**, 4992–5001 (1997).
- Cornali, E. *et al.* Vascular endothelial growth factor regulates angiogenesis and vascular permeability in kaposi's sarcoma. *Am. J. Pathol.* **149**, 1851–1869 (1996).
- Kyriakis, J. M. & Avruch, J. Sounding the alarm—Protein kinase cascades activated by stress and inflammation. *J. Biol. Chem.* **271**, 24313–24316 (1996).
- Samaniego, F., Markham, P. D., Gallo, R. C. & Ensoli, B. Inflammatory cytokines induce AIDS-Kaposi's sarcoma-derived spindle cells to produce and release basic fibroblast growth factor and enhance Kaposi's sarcoma-like lesion formation in nude mice. *J. Immunol.* **154**, 3582–3592 (1994).
- Gutkind, S., Novotny, E. A., Brann, M. R. & Robbins, K. C. Muscarinic acetylcholine receptors as agonist dependent oncogenes. *Proc. Natl Acad. Sci. USA* **88**, 4703–4707 (1991).
- Van Sande, J. *et al.* Genetic basis of endocrine disease: Somatic and germline mutations of the TSH receptor gene in thyroid diseases. *J. Clin. Endocrinol. Metab.* **80**, 2577–2585 (1995).
- Premont, R. T., Inglese, J. & Lefkowitz, R. J. Protein kinases that phosphorylate activated G-protein coupled receptors. *FASEB J.* **9**, 175–182 (1995).
- Geras Raaka, E. *et al.* Inhibition of constitutive signalling of Kaposi's sarcoma associated herpesvirus G protein coupled receptor by protein kinases in mammalian cells in culture. *J. Exp. Med.* (submitted).
- Koch, A. E. *et al.* Interleukin-8 as a macrophage mediator of angiogenesis. *Science* **258**, 1798–1801 (1992).
- Kolch, W., Martiny-Baron, G., Kieser, A. & Marme, D. Regulation of the expression of the VEGF/VPS and its receptors: role in tumor angiogenesis. *Breast Cancer Res. Treat.* **36**, 139–155 (1995).
- Goto, E., Goto, K., Weindel, K. & Folkman, J. Synergistic effects of vascular endothelial growth factor and basic fibroblast growth factor on the proliferation and cord formation of bovine capillary endothelial cells within collagen gels. *Lab. Invest.* **69**, 508–517 (1993).
- Montesano, R., Pepper, M. S. & Orci, L. Paracrine induction of angiogenesis *in vitro* by Swiss 3T3 fibroblasts. *J. Cell Sci.* **105**, 1013–1024 (1993).
- Rickinson, A. B. & Kieff, E. in *Fields Virology* 3rd edn (eds Fields, B. N., Knipe, D. M. & Howley, P. M.) 2397–2446 (Lippincott-Raven, Philadelphia, 1996).
- Mesri, E. A., Federoff, H. J. & Brownlee, M. Expression of vascular endothelial growth factor from a defective herpes simplex virus type 1 amplicon vector induces angiogenesis in mice. *Circ. Res.* **76**, 161–167 (1995).
- Coso, O. A. *et al.* The small GTP-binding proteins Rac1 and Cdc42 regulate the activity of the JNK/SAPK signaling pathway. *Cell* **81**, 1137–1146 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank G. Lam for HUVECs, R. Nador and O. Flore for sharing data and for comments, and A. Chadburn for help with histopathological characterization. Plasmids encoding GRK-2 and GRK-5 were a gift from R. J. Lefkowitz. This work was supported by NIH grants from NIAID to E.A.M., from NCI to E.C., and from NIDDK to M.C.G.

Correspondence and requests for materials should be addressed to E.A.M. (e-mail: eamesri@mail.med.cornell.edu).

Activating *Smoothened* mutations in sporadic basal-cell carcinoma

Jingwu Xie*, Maximilien Murone†, Shih-Ming Luoh†, Anne Ryan‡, Qimin Gu§, Chaohui Zhang†, Jeannette M. Bonifas*, Ching-Wan Lam||, Mary Hynes¶, Audrey Goddard§, Arnon Rosenthal¶, Ervin H. Epstein Jr* & Frederic J. de Sauvage†

* Department of Dermatology, San Francisco General Hospital, 1001 Potrero Street, Rm 269, Bldg 100, University of California, San Francisco, California 94110, USA

Departments of † Molecular Oncology, ‡ Pathology and § Molecular Biology and ¶ Neuroscience, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA

|| Department of Chemical Pathology, Prince of Wales Hospital, Shatin, Hong Kong

Basal-cell carcinomas (BCCs) are the commonest human cancer¹. Insight into their genesis came from identification of mutations in the *PATCHED* gene (*PTCH*) in patients with the basal-cell nevus syndrome, a hereditary disease characterized by multiple BCCs and by developmental abnormalities²⁻⁷. The binding of Sonic hedgehog (SHH) to its receptor, *PTCH*, is thought to prevent normal inhibition by *PTCH* of *Smoothened* (*SMO*), a seven-span

transmembrane protein^{8,9}. According to this model, the inhibition of *SMO* signalling is relieved following mutational inactivation of *PTCH* in basal-cell nevus syndrome. We report here the identification of activating somatic missense mutations in the *SMO* gene itself in sporadic BCCs from three patients. Mutant *SMO*, unlike wild type, can cooperate with adenovirus E1A to transform rat embryonic fibroblast cells in culture. Furthermore, skin abnormalities similar to BCCs developed in transgenic murine skin overexpressing mutant *SMO*. These findings support the role of *SMO* as a signalling component of the SHH-receptor complex and provide direct evidence that mutated *SMO* can function as an oncogene in BCCs.

Genomic DNA containing the human *SMO* locus was isolated in two non-overlapping λ clones spanning more than 35 kilobases (kb) and sequenced. The human gene is composed of 12 exons within 24 kb of genomic DNA (Fig. 1a). Exon 1 and 2 contain 5' untranslated sequences, the initiation codon ATG, and the entire signal peptide. The chromosomal location of human *SMO* was determined by fluorescence *in situ* hybridization. The λ clone *SMO*-7 was biotinylated and hybridized to metaphase spreads of a normal individual, resulting in specific labelling of the distal long arm of chromosome 7 (Fig. 1b). In addition, *SMO*-7 and a specific marker for the centromere of chromosome 7 (D7Z1) were co-hybridized, and measurement of 10 specifically labelled copies of chromosome 7 demonstrated that *SMO* is located 68% of the distance from the centromere to the telomere of chromosome arm 7q, an area corresponding to band 7q31-32.

As activation of the Hedgehog signalling pathway in BCCs and in other extracutaneous tumours^{10,11} can occur as a result of failure of inhibition by *PTCH* or because of activating changes in *SMO* gene. On the basis of the gene sequence, 13 primer pairs were synthesized and used to amplify each exon. Heteroduplex and single-strand conformation polymorphism (SSCP) analysis was performed on 47 sporadic BCCs. Variant bands were detected in exon 9 of two BCCs (BCC 913 and CBCC 9) and in exon 10 of one BCC (BCC 326) (Fig. 2a, b). No variant bands from exon 9 or 10 were seen in over 100 control DNA samples. Direct sequencing of exon 10 in BCC 326 identified a heterozygous missense mutation at base pair 1,685, designated M1, which is predicted to change

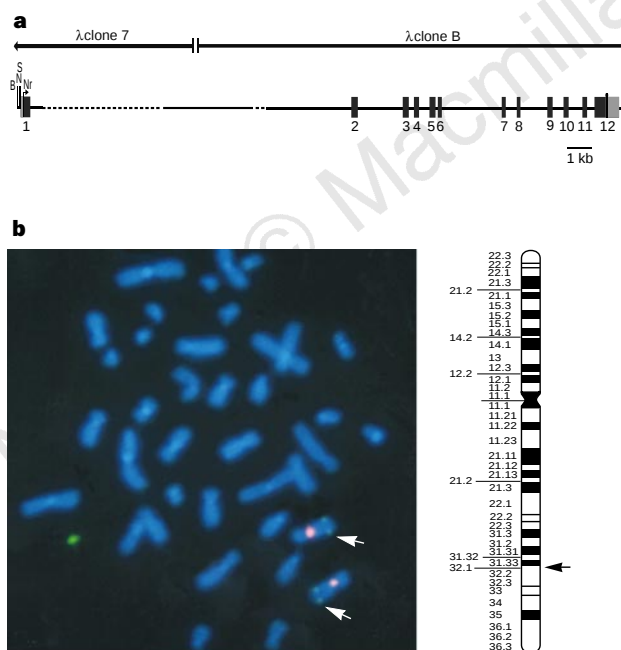


Figure 1 Human *SMO* genomic structure and localization. **a**, Genomic structure of *SMO*. Exons are represented by filled boxes; grey boxes represent non-coding regions and black boxes coding regions. The ATG initiation codon and the stop codon are indicated by lines topped by an arrow and a dot, respectively. The solid line indicates intronic regions for which sequence has been determined; dashed lines represent unsequenced gaps that have been sized by PCR. The CpG island upstream of exon 1 is shown as a cluster of rare restriction-enzyme recognition sites (B, *Bss*HII; S, *Sac*II; N, *Not*I; Nr, *Nar*I). The two non-overlapping λ clones characterized are shown above the gene structure. **b**, Chromosomal localization of the human *SMO* gene. Human metaphase cells were hybridized to a digoxigenin-labelled hSMO probe detected by a fluoresceinated antidigoxigenin antibody (green) and to a biotin-labelled centromeric marker of chromosome 7 (D7Z1) revealed by Texas-red avidin (red) and counterstained with DAPI. Of a total of 80 metaphase cells analysed, 71 were specifically labelled.

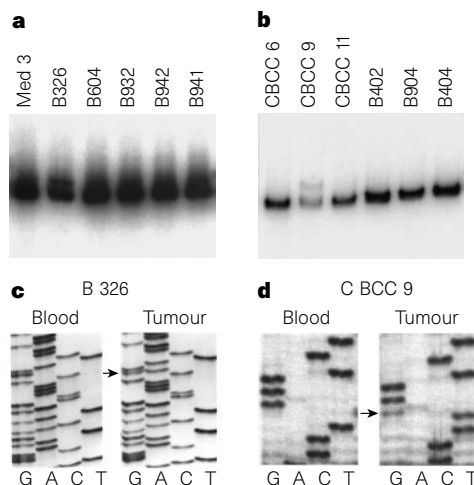


Figure 2 *SMO* mutations in sporadic BCC tumours. **a**, **b**, Heteroduplex analysis of sporadic BCC tumour DNA for exons 10 (**a**) and 9 (**b**). **c**, **d**, Sequencing of the heteroduplex variant PCR products shown in **a** and **b**, respectively, for patients B326 and C BCC9. The heterozygous mutations (arrows) are present in the tumour DNA and not in blood-cell DNA. The mutation in **d** encodes a Leu substitution at Trp 535; the mutation in **c** encodes a Gln substitution at Arg 562.

Arg 562 (CGG) to Gln (CAG) (Fig. 2c). DNA from the patient's blood did not contain this sequence change, indicating its somatic origin. This basic residue (either Arg or Lys) is conserved in *Drosophila*, rat and human. This BCC did not show loss of heterozygosity at 9q22.3 where *PTCH* is located, and no *PTCH* gene alterations were detected by SSCP.

Sequence analysis of the two BCCs with exon 9 heteroduplex variant bands identified the identical heterozygous mutation in each tumour, a G-to-T transversion at base pair 1,604 (M2), changing codon 535 from Trp to Leu (Fig. 2d). Again, patients' blood DNA samples did not contain this sequence alteration, indicating that the identical mutation arose in both tumours somatically. The Trp residue at codon 535 is conserved among all described SMO protein sequences, is located in the seventh transmembrane domain (Fig. 3), and corresponds to a conserved Tyr residue in most G-protein-coupled receptors. In particular, this residue in the α_{1B} -adrenergic receptor is thought to be involved in the formation of a polar pocket that keeps the receptor in a latent state. Disruption of this pocket by mutagenesis leads to constitutive signalling by the α_{1B} -adrenergic receptor¹⁴. The mutant Trp residue in SMO could lead to a similar result. Another mechanism by which these mutations could cause unrestrained SMO signalling would be by interference with the binding of SMO by PTCH, and hence a loss of the normal PTCH inhibition of SMO function. However, as described previously for wild-type SMO⁸, SMO-M1 and SMO-M2 can be immunoprecipitated with antibodies against epitope-tagged PTCH when

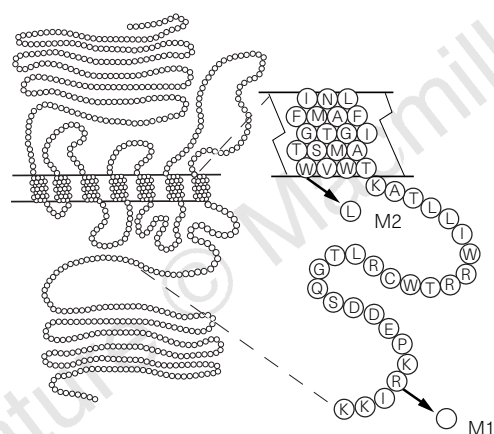


Figure 3 Predicted structure of the human SMO protein⁸. The location of two mutations identified in three BCC sporadic tumours is indicated. The mutation Trp 535 to Leu in the seventh transmembrane domain was identified in two separate tumours (M2). One mutation, Arg 562 to Gln, was identified in the carboxy-terminal cytoplasmic tail of SMO (M1).

Table 1 Transformation of Ref 52 cells

DNA transfected	Expt 1	Expt 2
E1A + vector	0	0, 0
E1A + SMO WT	0	0, 0
E1A + SMO M1	85	29, 13
E1A + SMO M2	14	15, 2
E1A + SMO frameshift	0	0, 0
E1A + N- <i>ras</i>	>300	>300
SV40 large T	>500	ND

Numbers shown are numbers of dense foci per 75 ml flask. ND, not determined.

expressed with PTCH in 293 cells (data not shown). Despite this retention of binding, the mutations may affect the affinity of the interaction, or there may be multiple contact points between PTCH and SMO, with mutation affecting a site that is crucial for inhibiting signalling.

To investigate the oncogenic potential of the two SMO missense mutations in BCCs, we transfected rat embryonic fibroblast REF52 cells with E1A together with wild-type or mutant SMO (M1 or M2), and assessed for focal cells overgrowth. Cells transfected with wild-type SMO and E1A formed a uniform monolayer in culture. In contrast, E1A plus either mutant SMO induced a transformed phenotype with focus formation over the confluent monolayer (Table 1). Cells within the foci lose normal density-dependent growth inhibition, resulting in increased cellular packing and loss of fibroblast morphology. Foci from cells transformed with mutant SMO took on average one week longer to form than those resulting from transformation with N-*ras* plus E1A or with SV40 large T antigen. Cells from pooled foci were able to grow in soft agar, whereas normal REF52 cells could not grow under those conditions. No foci were observed in cells transfected with mutant SMO alone or with E1A plus a frameshifted SMO construct, or with E1A plus an empty vector. Immunohistochemical staining using an antibody directed against a tag inserted at the amino terminus of the SMO constructs revealed that foci induced by mutant SMO have a high level of expression of SMO, indicating that mutant SMO expression correlates with cell overgrowth. Overexpression of wild-type SMO in this system is not sufficient to cause cellular transformation, probably because its activity can be controlled by endogenous PTCH. Alternatively, SMO shares sequence and structure similarities with the Wnt receptors, Frizzled, and the wild-type protein may require the binding of an as-yet unidentified ligand to be activated, whereas SMO mutations might lead to ligand-independent signalling. The mechanism of transformation by the SMO mutants remains to be determined; however, one known downstream effector of SHH signalling is the zinc-finger transcription factor GLI. The *GLI* gene is amplified in human glioblastomas¹⁵ and cooperates with E1A to transform primary fibroblasts¹⁶. Using

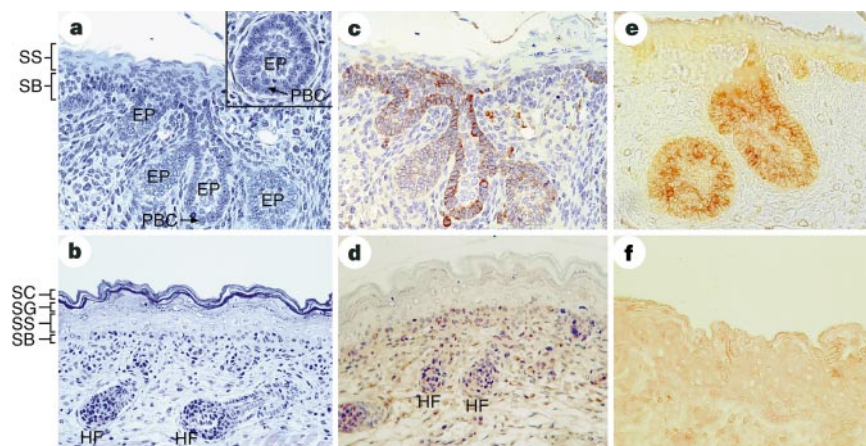


Figure 4 Skin of SMO-M2 transgenic mice display the features of basal-cell carcinoma. **a, b**, Haematoxylin-eosin-stained sections of SMO-M2 (**a**) and WT (**b**) transgenic E18 mouse skin. SB, stratum basale; SS, stratum spinosum; SC, stratum corneum; SG, stratum granulosum; PBC, palisading basal cells of epithelial Pegs (EP); HF, hair follicles. **c, d**, Overexpression of SMO detected in the epithelium of transgenic mice using an antibody directed against a tag inserted at the N terminus of the construct. Expression was limited to the basal cell layer in SMO-M2 mice (**c**) and not detected in wild-type mice (**d**). **e, f**, Accumulation of *PTCH* RNA, detected by *in situ* hybridization using a *PTCH* riboprobe in basal cells of mice overexpressing SMO-M2 (**e**) compared to wild-type mice (**f**).

real-time quantitative polymerase chain reaction (PCR)¹⁷, we have found that the levels of GLI messenger RNA were increased by around threefold in REF52 cell cultures established from pooled SMO-M1 and SMO-M2 foci (data not shown). Therefore, transformation by mutated SMO may be mediated in part by increased GLI activity.

To evaluate the effect of the SMO mutations *in vivo*, we used the keratin-5 (K5) promoter¹⁸ to drive expression of SMO-WT or SMO-M2 in the skin of transgenic mice. E18 embryos overexpressing SMO-WT did not display any morphological abnormalities compared to wild-type littermates. Although the macroscopic appearance of the skin in embryos overexpressing SMO-M2 was normal, microscopic examination of the skin demonstrated features consistent with those of human BCC (Fig. 4a, b). Unlike wild-type embryos, the surface epithelium was non-keratinized and keratohyalin granules were absent from the stratum granulosum. There was diffuse hyperplasia of the basal cell layer, which projected variable distances into the underlying dermis and formed irregular branching islands and ribbons of cells, bordered by palisading basal cells. Increased numbers of mitotic cells were present in the basal cell layers but there was no evidence of disruption of the basal membrane. This phenotype is reminiscent of the BCCs in mice overexpressing SHH¹², although these mice also develop other abnormalities such as fused digits and spina bifida. This disparity is consistent with the idea that SHH as an extracellular ligand can diffuse from the basal cell layer where it has been overexpressed and affect other tissue, whereas overexpression of SMO-M2, an intrinsic membrane protein, leads to a cell-autonomous defect, limited to basal cells. Expression of SMO-M2 under the control of the K5 promoter was found to be restricted to basal cells, as determined by staining of skin sections with an antibody directed against an epitope-tag (gD) inserted at the N terminus of the protein (Fig. 4c, d). Activation of the Hedgehog signalling pathway by overexpression of SMO-M2 was confirmed by the accumulation of large amounts of mRNA of *PTCH*, a target gene of SHH signalling, in basal cells (Fig. 4e, f).

Taken together, our results implicate the Hedgehog signalling pathway in tumorigenesis, especially in the skin. BCCs consistently show activation of the Hedgehog signalling pathway, as judged by increased *PTCH* mRNA^{6,19}. We have now provided evidence that this activation can be caused by missense mutations in SMO, supporting a model in which SMO activity drives the Hedgehog signalling pathway, the tumour suppressor *PTCH* represses signalling by SMO, and SHH relieves this repression. SMO can therefore be classified as a proto-oncogene. Although the SHH signalling pathway remains to be fully characterized, mutations of downstream targets of SMO may also be tumorigenic¹⁶. Pharmacological inhibition of SMO or of downstream effectors of this pathway could provide an effective treatment for BCCs and perhaps for other cancers as well. □

Methods

Isolation of the human SMO gene. Genomic clones were isolated by screening a human genomic library in lambda-Gem with human SMO cDNA. DNA from the λ clones was partially digested with *Sau3AI* or *Tsp50II* and fragments of 2–3 kb were subcloned in pGEM and sequenced by standard DYE-primer chemistry and analysed on an ABI377 automated sequencer. Gaps were covered by DYE-terminator sequencing of the appropriate PCR product.

Chromosomal localization. Mapping of human SMO was performed by fluorescent *in situ* hybridization to phytohaemagglutinin (PHA)-stimulated human lymphocyte chromosomes counterstained with propidium iodide and DAPI. Biotinylated probes were detected with avidin–fluorescein isothiocyanate (FITC). For two-colour experiments, the probe was detected by incubating the slides with fluoresceinated anti-digoxigenin antibodies and Texas-red avidin, followed by counterstaining with DAPI.

PCR and heteroduplex analysis. PCR amplification were performed as described³. MDE buffer was purchased from FMC and heteroduplex analysis

was carried out according to the manufacturer's protocol. Details of primers and of the intron/exon boundary sequences of SMO are available as Supplementary information. PCR product heteroduplex variants were sequenced directly using a Sequenase kit according to the protocol recommended by the manufacturer (United States Biochemical). A second independently amplified PCR product was sequenced to confirm the sequence change.

Focus-formation assay. REF52 cells were cotransfected with E1A and several different SMO constructs or with a vector control by using the calcium phosphate method^{20,21}. At 20 h post-transfection, cells were washed with PBS buffer and allowed to recover for 24 h in DMEM medium containing 10% fetal calf serum. The medium was changed once every two days. The formation of dense foci on top of the monolayer of untransformed cells was scored by direct counting of foci stained with 0.5% methylene blue at four to five weeks after transfection.

Transgenic mice. Constructs containing 6 kb of the human K5 promoter followed by an SV40 intron, human wild-type or M2 SMO cDNA and an SV40 polyadenylation site were injected into the pronucleus of fertilized C57B1/6 eggs. Genomic DNA was extracted from the tail and used to screen for integration of the transgene by PCR with the following hSMO primers: forward (5'-CGAGCTGAGAAGCGCCTGGGCC-3') and reverse (5'-GGTATTGGTT CCTCTCTTCTCTG-3').

Received 23 June; accepted 22 September 1997.

1. Miller, D. L. & Weinstock, M. A. Nonmelanoma skin cancer in the United States: incidence. *J. Am. Acad. Dermatol.* **30**, 774–778 (1994).
2. Hahn, H. et al. Mutations of the human homolog of *Drosophila patched* in the nevoid basal cell carcinoma syndrome. *Cell* **85**, 841–851 (1996).
3. Johnson, R. L. et al. Human homolog of *patched*, a candidate gene for the basal cell nevus syndrome. *Science* **272**, 1668–1671 (1996).
4. Chidambaram, A. et al. Mutations in the human homologue of the *Drosophila patched* gene in Caucasian and African-American nevoid basal cell carcinoma syndrome patients. *Cancer Res.* **56**, 4599–4601 (1996).
5. Uden, A. B. et al. Mutations in the human homologue of *Drosophila patched* (*PTCH*) in basal cell carcinomas and the Gorlin syndrome: different *in vivo* mechanisms of *PTCH* inactivation. *Cancer Res.* **56**, 4562–4565 (1996).
6. Gailani, M. R. et al. The role of the human homologue of *Drosophila patched* in sporadic basal cell carcinomas. *Nature Genet.* **14**, 78–81 (1996).
7. Wicking, C. et al. Most germ-line mutations in the nevoid basal cell carcinoma syndrome lead to a premature termination of the *PATCHED* protein, and no genotype–phenotype correlations are evident. *Am. J. Hum. Genet.* **60**, 21–26 (1997).
8. Stone, D. M. et al. The tumour-suppressor gene *patched* encodes a candidate receptor for Sonic hedgehog. *Nature* **384**, 129–134 (1996).
9. Marigo, V., Davey, R. A., Zuo, Y., Cunningham, J. M. & Tabin, C. J. Biochemical evidence that *patched* is the Hedgehog receptor. *Nature* **384**, 176–179 (1996).
10. Raffel, C. et al. Sporadic medulloblastomas contain *PTCH* mutations. *Cancer Res.* **57**, 842–845 (1997).
11. Xie, J. et al. Mutations of the *PATCHED* gene in several types of sporadic extracutaneous tumors. *Cancer Res.* **57**, 2369–2372 (1997).
12. Oro, A. E. et al. Basal cell carcinomas in mice overexpressing Sonic hedgehog. *Science* **276**, 817–821 (1997).
13. Fan, H., Oro, A. E., Scott, M. P. & Khavary, P. A. Induction of features of basal cell epithelioma in transgenic human skin expressing sonic hedgehog *in vivo*. *J. Invest. Dermatol.* **108**, 544 (1997).
14. Scheer, A., Fanelli, F., Costa, T., De Benedetti, P. G. & Cotecchia, S. The activation process of the alpha 1B-adrenergic receptor: potential role of protonation and hydrophobicity of a highly conserved aspartate. *Proc. Natl Acad. Sci. USA* **94**, 808–813 (1997).
15. Kinzler, K. W. et al. Identification of an amplified, highly expressed gene in a human glioma. *Science* **236**, 70–73 (1987).
16. Ruppert, J. M., Vogelstein, B. & Kinzler, K. W. The zinc finger protein GLI transforms primary cells in cooperation with adenovirus E1A. *Mol. Cell. Biol.* **11**, 1724–1728 (1991).
17. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res.* **6**, 986–994 (1996).
18. Byrne, C. & Fuchs, E. Probing keratinocyte and differentiation specificity of the human K5 promoter *in vitro* and in transgenic mice. *Mol. Cell. Biol.* **13**, 3176–3190 (1993).
19. Uden, A. B., Kerstin, B., Zaphiropoulos, P. G., Toftgard, R. & Backdal, M. S. The gene for Gorlin's syndrome, human *patched*, is consistently overexpressed in both hereditary and sporadic basal cell carcinomas. *J. Invest. Dermatol.* **108**, 596 (1997).
20. Wigler, M. et al. DNA-mediated transfer of the adenine phosphoribosyltransferase locus into mammalian cells. *Proc. Natl Acad. Sci. USA* **76**, 1373–1376 (1979).
21. Zhu, J., Rice, P. W., Gorsch, L., Abate, M. & Cole, C. N. Transformation of a continuous rat embryo fibroblast cell line requires three separate domains of simian virus 40 large T antigen. *J. Virol.* **66**, 2780–2791 (1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank D. Sherman and V. Bhagwandin for technical help, J. Ramirez for histology, K. Hagler for immunohistochemistry, E. Nelson for help with the animals, the Genentech DNA synthesis group, C. Hoffman and W. Anstine for preparing the figures, E. Fuchs for the K5 promoter, J. Gitscher for facilitating our collaboration and F. McCormick and M. Scott for advice and discussions. J.X. was the recipient of a Neutrogena fellowship of the Dermatology Foundation. Research in San Francisco was supported by the National Institute of Arthritis, Musculoskeletal and Skin Diseases.

Correspondence and requests for materials should be addressed to F.d.S. (e-mail: sauvage@gene.com) or E.H.E. (e-mail: chepestein@orca.ucsf.edu).

Induction of filopodium formation by a WASP-related actin-depolymerizing protein N-WASP

Hiroaki Miki, Takuya Sasaki*, Yoshimi Takai* & Tadaomi Takenawa

Department of Biochemistry, Institute of Medical Science, University of Tokyo, Shirokanedai 4-6-1, Minato-ku, Tokyo 108, Japan

* Department of Molecular Biology and Biochemistry, Osaka University Medical School, Suita, Osaka 565, Japan

Cdc42 is a small GTPase of the Rho family which regulates the formation of actin filaments to generate filopodia^{1,2}. Although there are several proteins such as PAK³, ACK⁴ and WASP (Wiskott-Aldrich syndrome protein)⁵ that bind Cdc42 directly, none of these can account for the filopodium formation induced by Cdc42. Here we demonstrate that before it can induce filopodium formation, Cdc42 must bind a WASP-related protein, N-WASP, that is richest in neural tissues⁶ but is expressed ubiquitously. N-WASP induces extremely long actin microspikes only when co-expressed with active Cdc42, whereas WASP, which is expressed in haematopoietic cells, does not, despite the structural similarities between WASP and N-WASP. In a cell-free system, addition of active Cdc42 significantly stimulates the actin-depolymerizing activity of N-WASP, creating free barbed ends from which actin polymerization can then take place. This activation seems to be caused by exposure of N-WASP's actin-depolymerizing region induced by Cdc42 binding.

To examine the functional interaction of N-WASP or WASP with Cdc42, we transiently co-expressed N-WASP (wild type or mutants) or WASP with a constitutively active mutant of Cdc42 (Myc-tagged Cdc42-G12V, which has valine substituted for glycine at residue 12) in COS 7 cells. The expression of active Cdc42 induced the formation of several actin microspikes (Fig. 1a, f). By contrast, wild-type N-WASP-expressing cells show no marked morphological changes and N-WASP was distributed mainly at the plasma membrane (data not shown). When N-WASP was co-expressed with active Cdc42, the formation of extremely long microspikes occurred in about half of the anti-Myc-positive cells (Fig. 1b, g). We next examined the effect of expression of WASP, an N-WASP homologue, on microspike formation induced by Cdc42. Ectopically expressed WASP localizes in the cytoplasm, especially in perinuclear areas (data not shown). Co-expression of WASP with active Cdc42 seems to suppress microspike formation by Cdc42 (Fig. 1c, h). This result is consistent with previous studies using PAE cells⁵ and suggests that WASP is not a downstream effector for filopodium formation.

As we reported previously, N-WASP directly depolymerizes actin filaments through its cofilin (cof)-homology region⁶. We therefore tested the effect of expressing a mutant Δ cof lacking four amino acids that are highly conserved in cofilin family members and found that Δ cof protein inhibited microspike formation by Cdc42 (Fig. 1d, i), suggesting that Δ cof functions in a dominant-negative manner. This result indicates that N-WASP is a critical effector of Cdc42 for microspike formation.

To confirm that these actin microspikes are caused by membrane protrusions (filopodia) but not by membrane retractions (retraction fibres), we used time-lapse microscopic analysis. Injection of active Cdc42 protein induced only small membrane protrusions in COS-7 cells (Fig. 1k): about 50% of the injected cells (59 out of 128 cells) showed such morphology. In contrast, co-injection with recombinant N-WASP protein caused much larger protrusions to

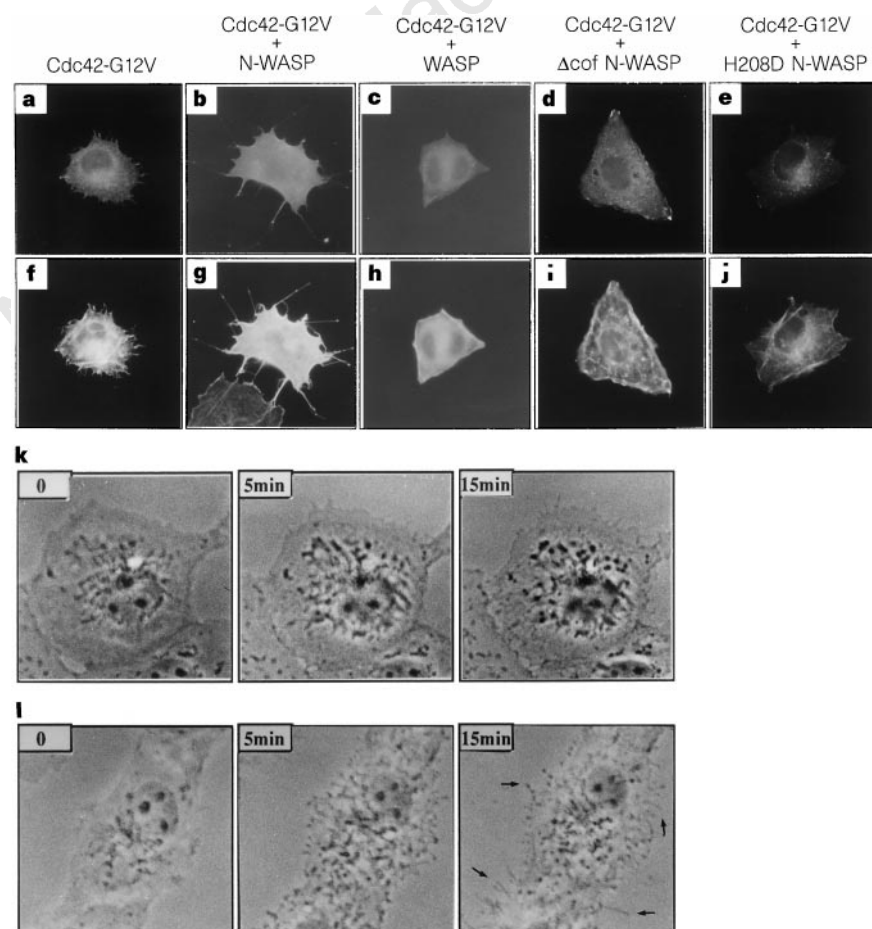


Figure 1 Enhancement of Cdc42-induced filopodium formation by N-WASP but not by WASP. COS-7 were transfected as indicated at the top. They were then stained with anti-Myc antibody (a-e) and phalloidin (f-j). In k and l, COS-7 cells were injected with 0.4 mg ml⁻¹ Cdc42-G12V mixed with (l) or without (k) 1 mg ml⁻¹ N-WASP. After injection, phase-contrast images were photographed at 5 and 15 min. Arrows in l indicate filopodia.

form (arrows in Fig. 1l). Formation of extremely long filopodia was seen in about 50% of cells (90 out of 184 cells) whereas more than 80% of cells (158 out of 184 cells) showed filopodium formation.

Next, we examined the interaction of N-WASP with Cdc42. First, Cdc42-binding proteins from brain cytosol were analysed by western blot with anti-N-WASP antibody (Rac-binding proteins were analysed as controls). As shown in Fig. 2a (top), N-WASP specifically associated with GTP-bound (active) Cdc42. The putative binding site is a GBD⁵/CRIB⁷ motif, through which WASP binds Cdc42^{5,8}. We therefore prepared a mutant (H208D) in which a histidine residue (H) conserved in all GBD/CRIB motifs is replaced by an aspartate (D) by site-directed mutagenesis. We then did a far-western blot analysis using Cdc42 as a probe against glutathione-S-transferase (GST) fusion proteins containing the GBD/CRIB motif of N-WASP (wild type or H208D mutant). As a result, [³²P]GTP-bound Cdc42 was found to associate specifically with GST-GBD/CRIB (Fig. 2b). Furthermore, this binding was abolished in the mutant with the H208D point mutation. These results indicate that there is a direct interaction between Cdc42 and N-WASP through the GBD/CRIB motif.

We did co-immunoprecipitation experiments to confirm that active Cdc42 also associates *in vivo* with wild-type N-WASP but not with the H208D mutant in COS-7 cells (data not shown), and to test the effect of this mutation on microspike formation by Cdc42. We found that this co-expression suppressed microspike formation by Cdc42 (Fig. 1e, j), indicating that Cdc42 must bind N-WASP for microspike formation to occur.

It has been reported⁹ that a Y40C mutant of Cdc42, which is unable to bind either PAK or WASP, can still induce filopodium

formation, suggesting that PAK and WASP are not important for filopodium formation. We therefore tested whether N-WASP can bind to Cdc42-Y40C using a far-western blot assay and found that Cdc42-Y40C binds to the N-WASP GBD/CRIB motif (Fig. 2c). By contrast, the Y40C mutant did not bind to a WASP GBD/CRIB motif fragment under the same conditions (only a very weak signal is seen after a much longer exposure; Fig. 2c, right, although wild-type Cdc42 binds to the same fragment).

Bradykinin treatment induces actin microspikes, including filopodia, in Swiss 3T3 cells in a process that is dependent on Cdc42 (ref. 2). As shown in Fig. 3a, western blot analysis using anti-N-WASP antibody reveals that Swiss 3T3 cells also contain N-WASP. We tested whether endogenous N-WASP participates in this process by microinjecting anti-N-WASP antibody into cells. Serum-starved cells were microinjected with anti-N-WASP antibody or control immunoglobulin G (IgG) and then treated with bradykinin. In control IgG-injected cells, bradykinin treatment induced morphological changes that reflected the formation of both filopodia and retraction fibres (Fig. 3b). No significant difference from the surrounding non-injected cells was observed (data not shown). However, anti-N-WASP antibody injection clearly suppresses the morphological changes and the injected cells take on a smooth shape normally seen in unstimulated cells (Fig. 3b). This result indicates that endogenous N-WASP is crucial for the actin cytoskeletal reorganization in Swiss 3T3 cells induced by bradykinin.

The carboxy-terminal fragment of N-WASP containing the verprolin- and cofilin-homologous regions and the highly acidic region (fragment VCA) has actin-depolymerizing activity, as shown by using recombinant proteins expressed in *Escherichia coli*⁶. To

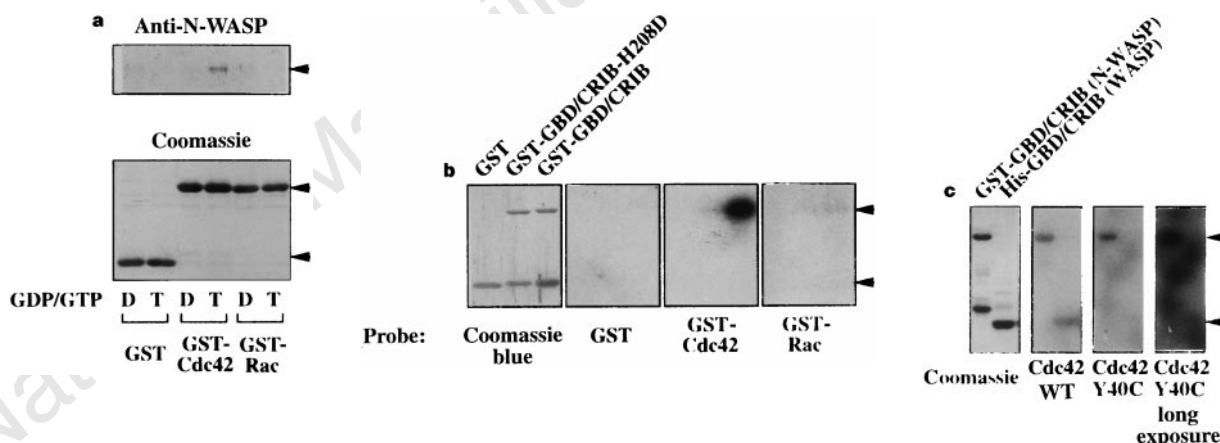


Figure 2 Specific and direct binding of N-WASP to activated Cdc42 through the GBD/CRIB motif. **a**, *In vitro* association of N-WASP with GTP-bound Cdc42. Immobilized GST-Cdc42 and Rac, loaded with GTP (T) or GDP (D), were mixed with rat brain lysates. Bound proteins were analysed by western blot with anti-N-WASP antibody. **b**, Direct binding of Cdc42 to the GBD/CRIB motif of N-WASP.

GST-GBD/CRIB (wild-type or H208D mutant) were probed with [³²P]GTP-bound GST-Cdc42 and Rac. **c**, Binding of Y40C mutant Cdc42 to N-WASP but not to WASP. GST-GBD/CRIB of N-WASP and histidine-tagged GBD/CRIB of WASP were tested for binding to wild-type and Y40C-mutant Cdc42 as for **b**. Right panel, longer exposure for the Y40C mutant.

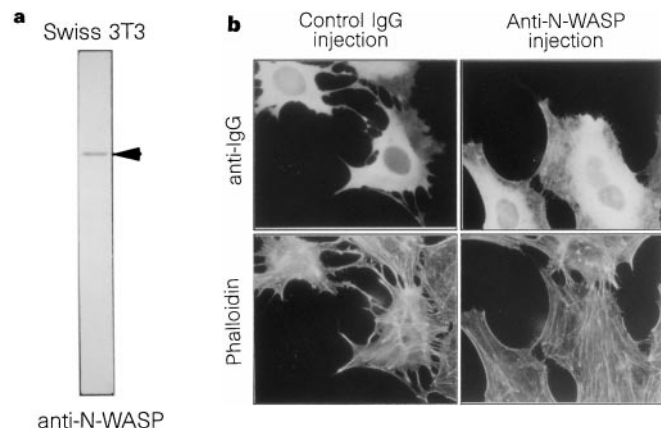


Figure 3 Requirement for endogenous N-WASP for microspike formation in Swiss 3T3 cells. **a**, Western blot analysis of Swiss 3T3 cell lysates using anti-N-WASP antibody; N-WASP is indicated (arrowhead). **b**, Inhibition of bradykinin-induced microspike formation by injection of anti-N-WASP antibody. Serum-starved Swiss 3T3 cells were microinjected with anti-N-WASP antibody (5 mg ml⁻¹) or control immunoglobulin G (5 mg ml⁻¹); 90 min after injection, cells were treated with 100 ng ml⁻¹ bradykinin for 10 min and then stained with anti-IgG antibody and phalloidin.

investigate whether full-length N-WASP also has actin-depolymerizing activity and, if so, how it is regulated, we prepared full-length N-WASP proteins (wild type and H208D mutant) using recombinant baculoviruses. Cell lysates were collected from baculovirus-infected Sf9 cells and the recombinant N-WASP proteins were purified to homogeneity on a heparin column (Fig. 4a). Using these preparations, we analysed their effect on actin filaments. Actin-depolymerizing proteins such as cofilin decrease the viscosity of actin filaments, reflecting the shortening of filament length¹⁰. A 0.1 molar ratio of cofilin to actin reduces the viscosity to the basal level¹¹. We therefore examined the effect of N-WASP on the viscosity of actin filaments using falling-ball methods. As shown in Fig. 4b, the addition of 0.5 μ M N-WASP to 5 μ M actin filaments lowered the viscosity slightly, but the effect was weaker than with cofilin; however, addition of GTP-loaded Cdc42 (1 μ M) to N-WASP significantly decreased the viscosity to the G-actin level. A weak effect of GTP-Cdc42 was also evident at 0.5 μ M, but GDP-Cdc42 had no significant effect. Addition of GTP- or GDP-Cdc42 alone (that is, without N-WASP) also had no effect. These results indicate that the actin-depolymerizing activity of N-WASP is regulated by direct binding of activated Cdc42 to N-WASP. In addition, GTP-Cdc42 had no effect on the H208D mutant, which does not bind Cdc42 (Fig. 4b).

We speculated that binding of Cdc42 to N-WASP might expose the VCA actin-depolymerizing region, which could be masked in unstimulated full-length N-WASP. The C-terminal region of N-WASP is rich in the acidic residues aspartate and glutamate⁶ and there is a very basic region just next to the N terminus of the GBD/CRIB motif⁶, so 'masking' may arise from an electrostatic association between these regions. We tested this by examining the binding of a histidine-tagged GBD/CRIB protein to GST-VCA protein immobilized covalently on agarose beads. As shown in Fig. 4c, the GBD/CRIB protein specifically associates with the VCA protein, whereas the histidine-tagged pleckstrin-homology (PH) domain of N-WASP, used as a control, does not bind the VCA protein (data not shown). This association is inhibited by addition of GTP-Cdc42 but not of GDP-Cdc42 (Fig. 4c), in support of our hypothesis.

We propose a model for N-WASP function in filopodium

formation. In the resting state, the actin-depolymerizing VCA region is masked by association with the GBD/CRIB region, keeping N-WASP inactive. Actin filaments are also stable, with most of the barbed ends capped. When some signal stimulates GDP/GTP exchange on Cdc42, the GTP-bound Cdc42 binds directly to N-WASP through the GBD/CRIB motif and exposes the VCA region. The actin filaments are partially depolymerized and the uncapped filament ends become exposed. At the uncapped filament ends, actin polymerization will occur. It has been suggested that profilin, an actin-binding protein, promotes actin polymerization at the uncapped barbed ends¹². In fact, N-WASP directly binds to profilin, probably through several polyproline stretches (data not shown). This actin polymerization induced by Cdc42-N-WASP could form the basis of filopodium formation. □

Methods

Transient expression in COS-7 cells. Rat N-WASP and human WASP cDNAs were used¹³. Mutagenesis (point mutation or deletion) was done by using PCR. Wild-type, Δ cof (lacking amino acids 473–476 (KRSK)), and H208D N-WASP and WASP cDNAs were ligated into pcDL-SR α plasmid⁶. Myc-tagged Cdc42-G12V and Rac-G12V in pEF-Bos plasmid were prepared as described¹⁴. Transfection was performed as before⁶.

Assay of binding of small G proteins with N-WASP. Immobilized GST-Cdc42 or Rac were loaded with GDP or GTP, then mixed with rat brain lysate at 4 °C for 2 h. After washing, beads were suspended in SDS-sample buffer. In far-western blot analysis, GST, GST-GBD/CRIB (amino acids 124–274 of wild-type and H208D mutant N-WASP), and histidine-tagged GBD/CRIB of WASP (amino acids 149–310) were applied to SDS-PAGE and transferred to polyvinylidene difluoride membranes. The membranes were incubated with [γ -³²P]GTP-loaded GST, GST-Cdc42, or GST-Rac (1 μ g ml⁻¹) for 5 min at room temperature, washed, and autoradiographed.

Microinjection. Antibodies (5 mg ml⁻¹) were injected into serum-starved Swiss 3T3 cells using an Eppendorf microinjector at 50 hPa hectopascals (hPa) for 0.2 s. For time-lapse analysis, serum-starved COS-7 cells were injected with 0.4 mg ml⁻¹ Cdc42-G12V (cleaved from GST by thrombin digestion¹⁵), mixed with or without 1 mg ml⁻¹ N-WASP (produced in the baculovirus system) at 50 hPa for 0.2 s, then photographed at 5 and 15 min after injection.

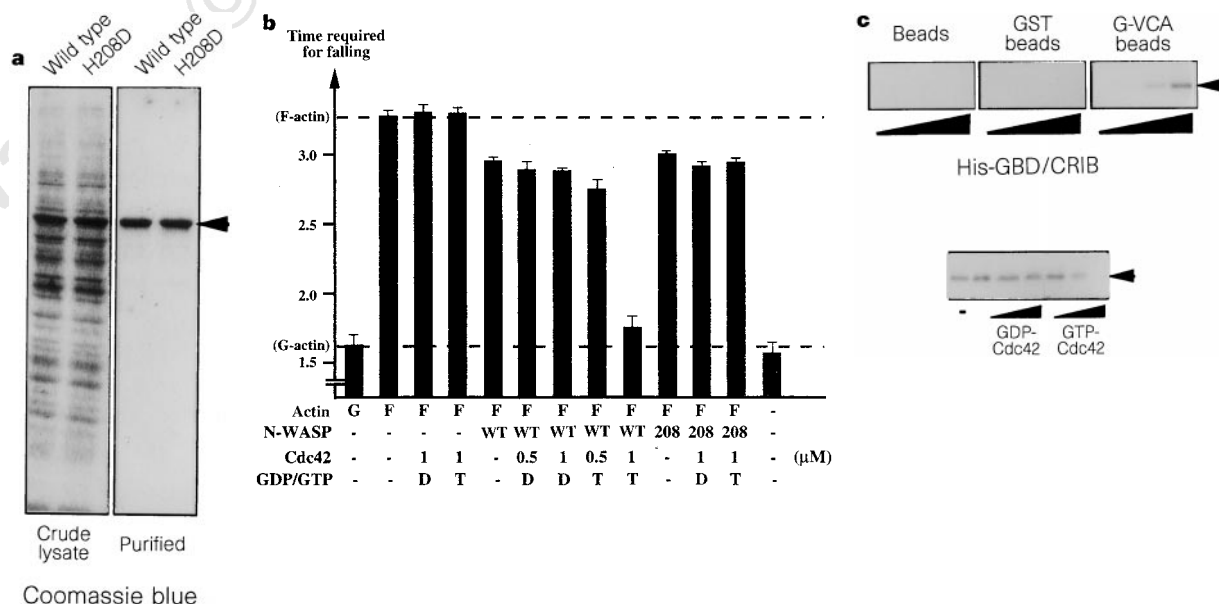


Figure 4 Effect of Cdc42 on depolymerization of actin filaments by N-WASP in a cell free system. **a**, Baculoviral expression and purification of recombinant N-WASP. Samples from the infected Sf9 cell lysates and the final purified fraction were checked by SDS-PAGE. **b**, Depolymerization of actin filaments by N-WASP is dependent on GTP-Cdc42. 5 μ M actin, 0.5 μ M N-WASP (wild-type (WT) or H208D mutant (208)), and GST-Cdc42 (GDP (D)- or GTP (T)-bound form) were mixed as indicated and polymerized. Viscosity was determined by the falling-ball

method; times required for falling are shown. Error bars represent standard deviations calculated from four independent experiments. **c**, Association between VCA and GBD/CRIB and the effect of Cdc42. His-tagged GBD/CRIB was applied to a GST-VCA column. Bound proteins were analysed by western blot with anti-His-tag antibody (top). Beads were further incubated with GDP- or GTP-loaded GST-Cdc42 (0.2, 1 or 5 μ M) and analysed as described (bottom).

Received 26 June; accepted 23 September 1997.

1. Nobes, C. D. & Hall, A. Rho, rac, and cdc GTPases regulate the assembly of multimolecular focal complexes associated with actin stress fibers, lamellipodia, and filopodia. *Cell* **81**, 53–62 (1995).
2. Kozma, R., Ahmed, S., Best, A. & Lim, L. The GTPase-activating protein n-chimaerin cooperates with Rac1 and Cdc42Hs to induce the formation of lamellipodia and filopodia. *Mol. Cell. Biol.* **15**, 1942–1952 (1995).
3. Manser, E., Leung, T., Salihuddin, H., Zhao, Z. S. & Lim, L. A brain serine/threonine protein kinase activated by Cdc42 and Rac1. *Nature* **367**, 40–46 (1994).
4. Manser, E., Leung, T., Salihuddin, H., Zhao, Z. S. & Lim, L. A non-receptor tyrosine kinase that inhibits the GTPase activity of p21^{cdc42}. *Nature* **363**, 364–367 (1993).
5. Symons, M. *et al.* Wiskott–Aldrich syndrome protein, a novel effector for the GTPase CDC42Hs, is implicated in actin polymerization. *Cell* **84**, 723–734 (1996).
6. Miki, H., Miura, K. & Takenawa, T. N-WASP, a novel actin-depolymerizing protein, regulates the cortical cytoskeletal rearrangement in a PIP2-dependent manner downstream of tyrosine kinases. *EMBO J.* **15**, 5326–5335 (1996).
7. Burbelo, P. D., Drechsel, D. & Hall, A. A conserved binding motif defines numerous candidate target proteins for both Cdc42 and Rac GTPases. *J. Biol. Chem.* **270**, 29071–29074 (1995).
8. Aspenstrom, P., Lindberg, U. & Hall, A. Two GTPases, Cdc42 and Rac, bind directly to a protein implicated in the immunodeficiency disorder Wiskott–Aldrich syndrome. *Curr. Biol.* **15**, 5725–5731 (1996).
9. Lamarche, N. *et al.* Rac and Cdc42 induce actin polymerization and G1 cell cycle progression independently of p65PAK and the JNK/SAPK MAP kinase cascade. *Cell* **87**, 519–529 (1996).
10. Nishida, E., Maekawa, S. & Sakai, H. Cofilin, a protein in porcine brain that binds to actin filaments and inhibits their interactions with myosin and tropomyosin. *Biochemistry* **23**, 5307–5317 (1984).
11. Aizawa, H. *et al.* Identification, characterization, and intracellular distribution of cofilin in *Dictyostelium discoideum*. *J. Biol. Chem.* **270**, 10923–10932 (1995).
12. Pantaloni, D. & Carlier, M.-F. How profilin promotes actin filament assembly in the presence of thymosin β 4. *Cell* **75**, 1007–1014 (1993).
13. Fukuoaka, M., Miki, H. & Takenawa, T. Identification of N-WASP homologs in human and rat brain. *Gene* **196**, 43–48 (1997).
14. Komuro, R., Sasaki, T., Takaishi, K., Orita, S. & Takai, Y. Involvement of Rho and Rac small G proteins and Rho GDI in Ca²⁺-dependent exocytosis from PC12 cells. *Genes to Cells* **1**, 943–951 (1996).
15. Self, A. J. & Hall, A. Purification of recombinant Rho/Rac/G25K from *Escherichia coli*. *Meth. Enzymol.* **256**, 3–10 (1995).

Acknowledgements. We thank S. Suetsugu and Y. Banzai for their skilful technical assistance. This work was supported in part by a Grant-in-Aid for Cancer Research from the Ministry of Education, Science, and Culture of Japan and a Grant-in-Aid for Research for the Future Program from the Japan Society for the Promotion of Science.

Correspondence and requests for materials should be addressed to T.T. (e-mail: takenawa@hgc.ims.u-tokyo.ac.jp).

Cleavage of CAD inhibitor in CAD activation and DNA degradation during apoptosis

Hideki Sakahira*, Masato Enari* & Shigekazu Nagata*†

* Department of Genetics, Osaka University Medical School, 2-2 Yamada-oka, Suita, Osaka 565, Japan

† Osaka Bioscience Institute, 6-2-4 Furuedai, Suita, Osaka 565, Japan

Various molecules such as cytokines and anticancer drugs, as well as factor deprivation, rapidly induce apoptosis (programmed cell death)^{1,2}, which is morphologically characterized by cell shrinkage and the blebbing of plasma membranes and by nuclear condensation^{3,4}. Caspases, particularly caspase 3, are proteases that are activated during apoptosis and which cleave substrates such as poly(ADP-ribose) polymerase, actin, fodrin, and lamin^{5,6}. Apoptosis is also accompanied by the internucleosomal degradation of chromosomal DNA^{7–9}. In the accompanying Article¹⁰, we have identified and molecularly cloned a caspase-activated deoxyribonuclease (CAD) and its inhibitor (ICAD). Here we show that caspase 3 cleaves ICAD and inactivates its CAD-inhibitory effect. We identified two caspase-3 cleavage sites in ICAD by site-directed mutagenesis. When human Jurkat cells were transformed with ICAD-expressing plasmid, occupation of the receptor Fas, which normally triggers apoptosis, did not result in DNA degradation. The ICAD transformants were also resistant to staurosporine-induced DNA degradation, although staurosporine still killed the cells by activating caspase. Our results indicate that activation of CAD downstream of the caspase cascade is responsible for internucleosomal DNA degradation during apoptosis, and that ICAD works as an inhibitor of this process.

We have shown that the growing cells carry an inactive form of CAD and its inhibitor ICAD in the cytosol. Mouse ICAD purified

from mouse WR19L cells is a protein of relative molecular mass (M_r) 32K, which can inhibit CAD-induced DNA degradation in nuclei as well as its DNase activity on plasmid DNA. In contrast to growing cells, lysates from Fas-activated apoptotic cells do not have ICAD activity (data not shown), suggesting that ICAD could be inactivated by an apoptotic signal. A candidate in activator is caspase 3, which is itself activated during apoptosis^{1,11–13}. To test whether ICAD is cleaved by caspase 3, we incubated ICAD purified from mouse WR19L cells with caspase 3. As shown in Fig. 1a, caspase-3-treated ICAD could no longer inhibit CAD-induced DNA degradation in nuclei or its DNase activity on plasmid DNA; also, the 32K ICAD protein was specifically cleaved by caspase 3 (Fig. 1b).

Molecular cloning of ICAD complementary DNAs has indicated that ICAD exists in two forms, ICAD-L for the long form and ICAD-S for the short form¹⁰, which seem to be generated by alternative splicing. Both ICAD-L and ICAD-S carry two putative

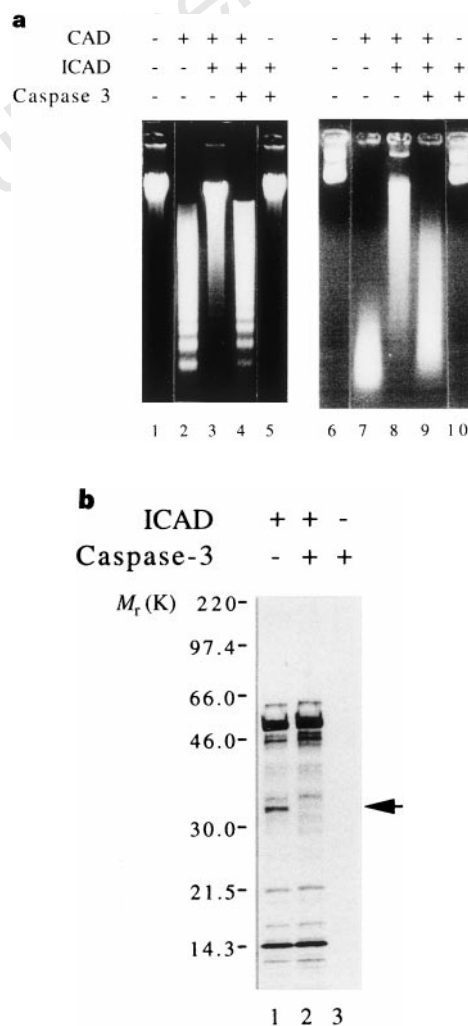


Figure 1 Inactivation of ICAD by caspase-3 cleavage. **a**, Native purified ICAD¹⁰ (100 ng) was incubated with 15 ng caspase 3 in buffer A (see Methods). Lysates from Fas-activated W4 cells (20 μ g) (lanes 1–5) or caspase-3-activated, partially purified CAD (5 ng) (lanes 6–10) were mixed with 25 ng caspase-3-treated or untreated ICAD. CAD activity was assayed with mouse nuclei (lanes 1–5) or with plasmid DNA (lanes 6–10) using CAD (lanes 2 and 7), CAD and ICAD (lanes 3 and 8), CAD and caspase-3-treated ICAD (lanes 4 and 9), or caspase-3-treated ICAD alone (lanes 5 and 10). **b**, Purified ICAD (100 ng) was treated with 15 ng caspase 3, electrophoresed on a 10–20% gradient polyacrylamide gel and visualized by silver staining. Lanes: 1, ICAD before caspase 3 treatment; 2, ICAD after caspase 3 treatment; 3, 15 ng caspase 3. ICAD is indicated by an arrow.

cleavage sites for caspase 3 (Fig. 2a); specifically, DEPD and DAVD at amino-acid positions 117 and 224 conform well to the consensus cleavage sites for caspase (refs 3, 14, 15). Caspase requires aspartic acid at the P1 position of the cleavage site^{14,15}. To test whether caspase 3 cleaves ICAD at these positions (Fig. 2a), the aspartic acid residues at the cleavage sites were singly or doubly mutated to glutamic acid. Recombinant glutathione-S-transferase (GST) fusion protein for each ICAD mutant was produced in *Escherichia coli* and purified to homogeneity. Treatment of wild-type ICAD-L (L/WT) with caspase 3 yielded two bands of M_r 40K and 12K (Fig. 2b). The upper band is a fusion protein consisting of GST and the amino-terminal portion of ICAD, whereas the lower band is actually two bands derived from the middle and carboxy-terminal portions of ICAD. When Asp 117 was mutated (L/D117E), the

upper band generated by caspase 3 shifted to an M_r of 52K, whereas mutation of Asp 224 (L/D224E) shifted the lower band to an M_r of 26K. Furthermore, the double mutation at both sites (L/dm) generated an ICAD that was resistant to caspase 3. Similar results were obtained with mutants of ICAD-S. We then assayed for inhibition by mutated ICAD of CAD-induced DNA degradation. Like native ICAD, caspase 3 treatment of recombinant ICAD destroyed its inhibitory activity (Fig. 2c). ICADs carrying a single mutation at Asp 117 or double mutations at Asp 117 and 224 were still functional after treatment with caspase 3. On the other hand, ICAD, particularly ICAD-S, carrying the intact cleavage site at Asp 117 was sensitive to caspase 3. These results indicate that the cleavage of ICAD at Asp 117 by caspase 3 inactivates ICAD, and that mutation at Asp 117 can generate caspase-3-resistant ICAD.

Human Jurkat cells can be killed by agonistic anti-Fas antibody or by a high concentration of staurosporine, and in both cases cell death is accompanied by internucleosomal cleavage of chromosomal DNA¹⁶ (A. Kawahara and S.N., unpublished results). To investigate the effect of ICAD on DNA degradation during apoptosis, wild-type ICAD-L and ICAD-S, as well as their caspase-3-resistant double mutants, were tagged with the Flag epitope and their expression plasmids introduced into Jurkat cells. The stable transformant clones expressing ICAD were selected by western blotting using anti-Flag antibody (Fig. 3a). The expression of Fas in all transformant clones was the same as in parental cells (data not shown). When parental Jurkat cells were treated with anti-Fas antibody, their chromosomal DNA was degraded into 180-base-pair (bp) multimers within 4 h (Fig. 3b). On the other hand, transformants expressing wild-type ICAD-L or ICAD-S, or their caspase-3-resistant mutants, did not show any DNA fragmentation, even after a 6-h incubation with anti-Fas antibody. Similarly, when Jurkat cells were incubated with 10 μ M staurosporine, their chromosomal DNA was degraded within 2 h; however, this degradation was completely inhibited by overexpression of wild-type ICAD-L or ICAD-S. These results indicate that ICAD specifically blocks the DNase(s) that is activated during apoptosis.

The killing of cells by staurosporine was then followed by flow cytometry using annexin V¹⁷. As shown in Fig. 3c, most of the parental Jurkat cells, as well as the transformants expressing wild-type ICAD-L or ICAD-S, became annexin-V-positive (indicative of cell death) within 2 h of treatment with staurosporine. When cell death was assayed by the MTT method, which determines the integrity of mitochondria¹⁸, the results also indicated that the cells (parental Jurkat and ICAD transformants) treated with staurosporine died within 2 h (data not shown). This killing was accompanied by the cleavage of 116K poly(ADP-ribose) polymerase to an 85K protein (Fig. 3d). Furthermore, when caspase 3 activity in the cytosol was determined by using a fluorescent substrate, the activity in ICAD transformant cells after staurosporine treatment was comparable to that in parental Jurkat cells (Fig. 3e). The anti-Fas antibody also killed the Jurkat and ICAD transformants by activating caspase 3, although the killing process was slow compared with that induced by staurosporine (data not shown). These results indicate that the cell can die without DNA degradation, probably as a result of cleavage of various important cellular substrates by caspases.

We have shown that ICAD can be inactivated by caspase 3 and that overexpression of ICAD in human Jurkat cells blocks DNA degradation during apoptosis. Mouse ICAD-L shares significant sequence similarity with one of the subunits of human DFF (DFF45, for DNA-fragmentation factor 45)¹⁹ and seems to be the murine counterpart of DFF45. Purified human DFF consists of 45K and 40K subunits and causes DNA fragmentation in nuclei when treated with caspase 3. This is also a property of the inactive CAD that we have identified¹⁰. DFF itself does not have DNase activity, suggesting that DFF is not human CAD. It has been proposed that when DFF45 is cleaved by caspase 3, DFF enters the nucleus and activates a

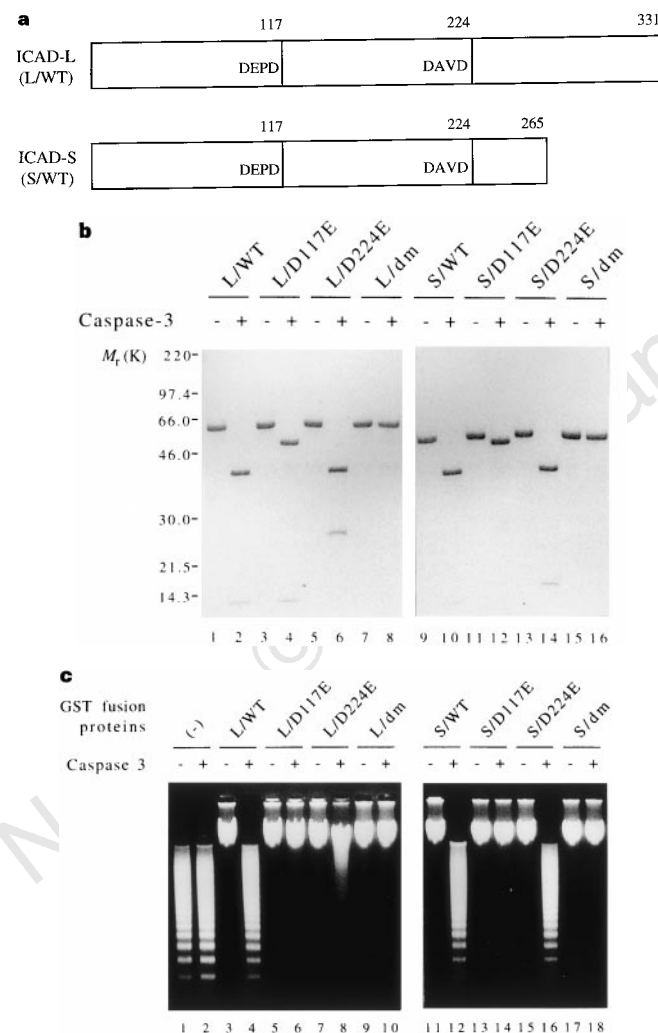


Figure 2 Identification of caspase 3 cleavage sites in ICAD. Recombinant GST fusion proteins with ICAD-L and ICAD-S, and their mutants, were produced in *E. coli* and purified as described in Methods. **a**, Putative caspase 3 cleavage sites in ICAD-L and ICAD-S. **b**, Recombinant GST fusion proteins (2 μ g) with ICAD-L (lanes 1–8) or ICAD-S (lanes 9–16) were incubated with 75 ng caspase 3. Samples were run on a 10–20% gradient polyacrylamide gel before (lanes 1, 3, 5, 7, 9, 11, 13 and 15) and after (lanes 2, 4, 6, 8, 10, 12, 14 and 16) caspase 3 treatment, and stained with Coomassie brilliant blue. The GST-ICAD proteins were: wild-type (L/WT and S/WT), singly mutated at D117 (L/D117E and S/D117E) or at D224 (L/D224E and S/D224E), or doubly mutated at D117 and D224 (L/dm and S/dm). **c**, 10 ng GST-ICAD-L (lanes 3–10) or GST-ICAD-S (lanes 11–18) were incubated with (lanes 2, 4, 6, 8, 10, 12, 14, 16 and 18) or without (lanes 1, 3, 5, 7, 9, 11, 13, 15 and 17) 15 ng caspase 3 as described in Methods. Cell lysate from Fas-activated W4 cells (20 μ g) was then added and CAD activity assayed on liver nuclei.

nuclear DNase¹⁹. In contrast, mouse CAD has an intrinsic DNase activity, which ICAD inhibits by binding to it. These results may indicate that ICAD can bind not only CAD, but also another molecule (DFF40) that regulates a DNase. The DNase regulated by DFF may be CAD or a CAD-like DNase because overexpression of ICAD inhibits DNA degradation in the apoptotic process in human cells.

Caspase 3 inactivates ICAD by cleaving it. We originally thought that CAD could be inhibited in cells only by caspase-resistant ICAD: however, overexpression of wild-type ICAD in Jurkat cells completely blocks DNA degradation induced by Fas engagement, as well as by staurosporine. These results suggest that an ICAD-inhibitable DNase is activated during apoptosis and that this activation can be triggered by different stimuli. As ICAD inhibits CAD, but does not inhibit DNase I or II (ref. 10), we conclude that CAD or a CAD-like DNase is responsible for the DNA degradation that occurs in apoptosis, at least in human Jurkat cells where apoptosis is induced by Fas or staurosporine. Most ICAD introduced into the cells is quickly degraded by these apoptotic stimuli (H.S. and S.N., unpublished results), indicating that a small amount of ICAD is sufficient to inhibit CAD activity, or that the concentration of CAD is minimal. Apoptosis can be induced in thymocytes by glucocorticoids⁹ and in various cells by factor or serum deprivation². To

assess the generality of CAD's involvement in apoptosis, it will be necessary to study the effect of ICAD in other systems.

Studies on programmed cell death in *C. elegans* have indicated that DNase activity is not essential for cell killing²⁰. One mouse cell line was killed by the action of cytotoxic lymphocytes without apparent DNA fragmentation²¹. Here the transformants expressing ICAD were killed by staurosporine treatment without cleavage of chromosomal DNA. Caspase was fully active under these conditions, which places the apoptotic DNase downstream of the caspases and indicates that cleavage of cellular substrates such as actin, Rho-GDI and fodrin is sufficient to kill the cells^{5,6}. In addition to cleavage of DNA, apoptotic stimuli induce the condensation and fragmentation of nuclei³. It is not yet clear whether it is DNA degradation or cleavage of nuclear proteins such as poly(ADP-ribose) polymerase and lamin that is responsible for the morphological changes in cells^{4,22,23}; cells that carry a fully activated caspase and do not suffer DNA degradation will clarify this point. Finally, it has been proposed that degradation of chromosomal DNA is involved in the cleaning up of dead cells, by facilitating phagocytosis or by preventing transformation of the phagocytosing cells by intact DNA from the dying cells²². The genes encoding CAD and ICAD will be useful for addressing these questions. □

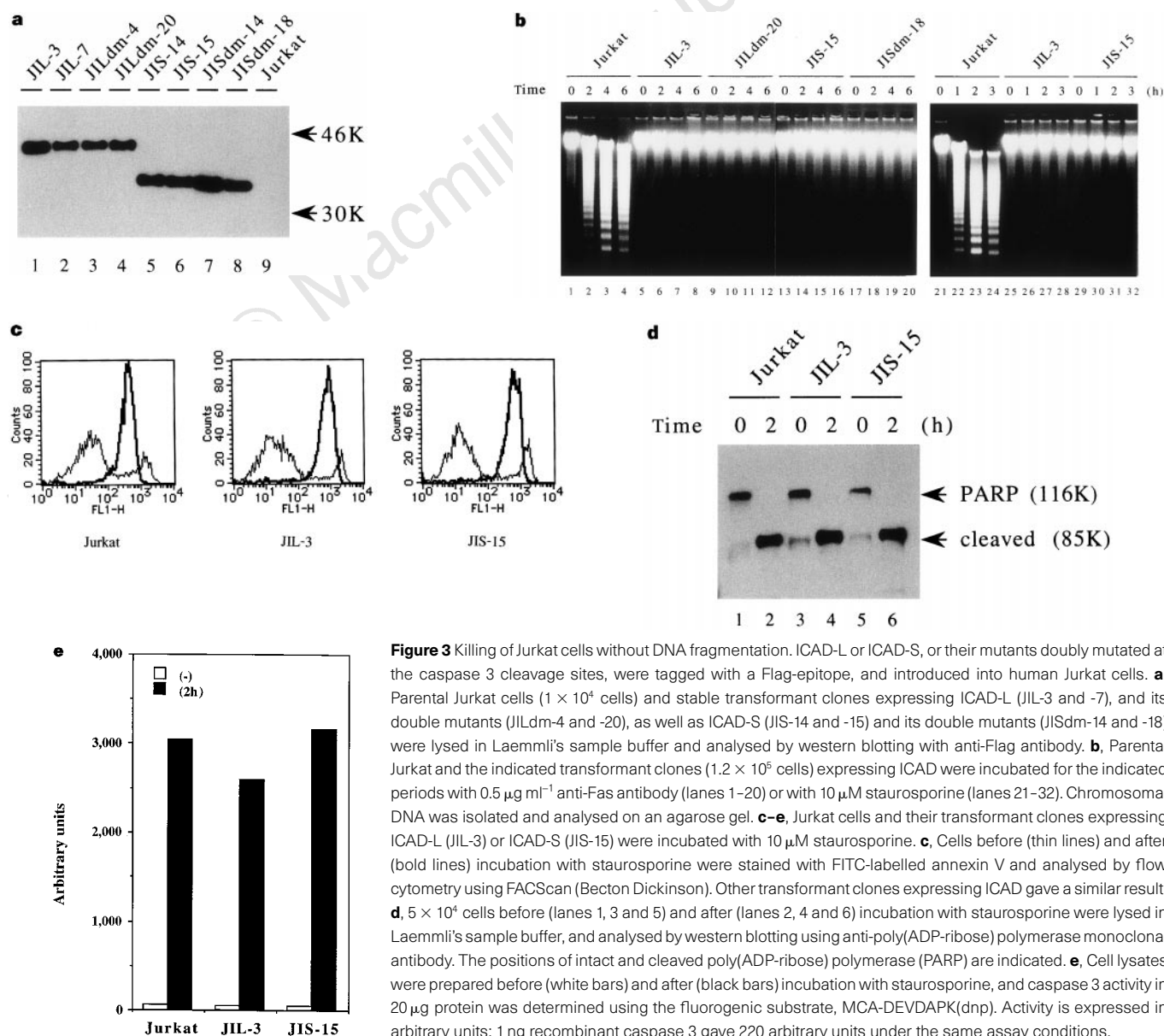


Figure 3 Killing of Jurkat cells without DNA fragmentation. ICAD-L or ICAD-S, or their mutants doubly mutated at the caspase 3 cleavage sites, were tagged with a Flag-epitope, and introduced into human Jurkat cells. **a**, Parental Jurkat cells (1×10^4 cells) and stable transformant clones expressing ICAD-L (JIL-3 and -7), and its double mutants (JILdm-4 and -20), as well as ICAD-S (JIS-14 and -15) and its double mutants (JISdm-14 and -18) were lysed in Laemmli's sample buffer and analysed by western blotting with anti-Flag antibody. **b**, Parental Jurkat and the indicated transformant clones (1.2×10^5 cells) expressing ICAD were incubated for the indicated periods with 0.5 $\mu\text{g ml}^{-1}$ anti-Fas antibody (lanes 1–20) or with 10 μM staurosporine (lanes 21–32). Chromosomal DNA was isolated and analysed on an agarose gel. **c–e**, Jurkat cells and their transformant clones expressing ICAD-L (JIL-3) or ICAD-S (JIS-15) were incubated with 10 μM staurosporine. **c**, Cells before (thin lines) and after (bold lines) incubation with staurosporine were stained with FITC-labelled annexin V and analysed by flow cytometry using FACScan (Becton Dickinson). Other transformant clones expressing ICAD gave a similar result. **d**, 5×10^4 cells before (lanes 1, 3 and 5) and after (lanes 2, 4 and 6) incubation with staurosporine were lysed in Laemmli's sample buffer, and analysed by western blotting using anti-poly(ADP-ribose) polymerase monoclonal antibody. The positions of intact and cleaved poly(ADP-ribose) polymerase (PARP) are indicated. **e**, Cell lysates were prepared before (white bars) and after (black bars) incubation with staurosporine, and caspase 3 activity in 20 μg protein was determined using the fluorogenic substrate, MCA-DEVDAPK(dnp). Activity is expressed in arbitrary units; 1 ng recombinant caspase 3 gave 220 arbitrary units under the same assay conditions.

Methods

Purified ICAD, and assay for CAD and ICAD. Purification of ICAD from mouse WR19L cells has been described¹⁰. The final preparation (active fractions from the Superdex-200 gel-filtration column) was heated at 90 °C for 15 min and centrifuged at 100,000g for 30 min; the supernatant was used as purified ICAD. CAD activity was determined by *in vitro* assay for apoptosis using mouse liver nuclei, or by measuring DNase activity using plasmid DNA as substrate¹⁰. ICAD activity was assayed by inhibition of CAD activity. The cytosolic fraction from Fas-activated cells was used as a source of activated CAD¹⁰. In some cases, partially purified CAD (100 ng) (resource-S fraction¹⁰) was incubated with 150 ng caspase 3 at 4 °C for 12 h and at 30 °C for 1 h in 100 µl buffer A (10 mM HEPES-KOH, pH 7.0, 5 mM MgCl₂, 5 mM EGTA, 10% glycerol, 1 mM DTT) containing 50 mM NaCl and 100 µg ml⁻¹ BSA, and used as activated CAD after inactivating caspase 3 with Ac-DEVD-cho.

Expression of ICAD in *E. coli*. To generate the D117E and D224E mutants, ICAD-S and I-CAD-L cDNAs were mutated by recombinant PCR²⁴ using 20-nucleotide primers carrying the mutated nucleotides. The D117E/D224E double mutants were generated by replacing the N-terminal half of the D224E mutant with that of D117E mutant, using the *Bgl*/II site between D117 and D224. The wild-type and mutant cDNAs were fused to the GST gene of pGEX-2T[128/129]. pGEX-2T[128/129] is derived from pGEX-2T (Pharmacia) by in-frame insertion of the Flag-epitope tag and a heart-muscle kinase-recognition sequence following the GST gene²⁵. The fusion proteins were expressed in *E. coli* AD202, and purified by glutathione-Sepharose-4B (Pharmacia) according to the manufacturer's instructions. Recombinant ICAD eluted from the glutathione column was purified on a Mono-Q HR5/5 column with a NaCl gradient in buffer A containing 0.02% Tween-20.

Transformation of Jurkat cells, cell-death assay and western blotting. Wild-type ICAD-S and ICAD-L and their double mutants were tagged with the Flag epitope at their N termini and ligated into the pEF-BC3 vector²⁶. Human Jurkat cells were co-transfected with ICAD expression plasmids and pBLIhyg by electroporation²⁷. Hygromycin B (800 µg ml⁻¹)-resistant transformants were picked and expanded. Expression of ICAD in individual clones was analysed by western blotting with anti-Flag monoclonal antibody (clone M2, Kodak). To bring about apoptosis, Jurkat and ICAD-transformant cells (2.5 × 10⁵ cells per ml) were incubated at 37 °C with 0.5 µg ml⁻¹ of mouse anti-human Fas monoclonal antibody (clone CH11, MBL) or 10 µM staurosporine. Cell death was assayed by using the annexin V method¹⁷ (R & D System kit). DNA fragmentation was assayed as described²⁷. Proteolysis of poly(ADP-ribose) polymerase was determined by western blotting using a mouse monoclonal antibody against human poly(ADP-ribose) polymerase (clone C-2-10)²⁸. To determine cytosolic caspase 3 activity, a peptide (DEVDPK) carrying the caspase-3-recognition sequence was coupled to the highly fluorescent (7-methoxycoumarin-4-yl) acetyl group (MCA) and its quenching 2,4-dinitrophenyl (Dnp) group, and the product MCA-DEVDPK(dnp) used as a substrate as described¹¹.

Received 20 November; accepted 1 December 1997.

- Nagata, S. Apoptosis by death factor. *Cell* **88**, 355–365 (1997).
- Raff, M. C. Social controls on cell survival and cell death. *Nature* **356**, 397–400 (1992).
- Wyllie, A. H., Kerr, J. F. R. & Currie, A. R. Cell death: the significance of apoptosis. *Int. Rev. Cytol.* **68**, 251–306 (1980).
- Earnshaw, W. C. Nuclear changes in apoptosis. *Curr. Opin. Cell Biol.* **7**, 337–343 (1995).
- Henkart, P. A. ICE family protease: mediators of all apoptotic cell death? *Immunity* **4**, 195–201 (1996).
- Martin, S. & Green, D. Protease activation during apoptosis: death by a thousand cuts. *Cell* **82**, 349–352 (1995).
- Compton, M. M. A biochemical hallmark of apoptosis: internucleosomal degradation of the genome. *Cancer Metast. Rev.* **11**, 105–119 (1992).
- Wyllie, A. H., Morris, R. G., Smith, A. L. & Dunlop, D. Chromatin cleavage in apoptosis: association with condensed chromatin morphology and dependence on macromolecular synthesis. *J. Pathol.* **142**, 66–77 (1984).
- Wyllie, A. H. Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation. *Nature* **284**, 555–556 (1980).
- Enari, M. *et al.* A caspase-activated DNase that degrades DNA during apoptosis, and its inhibitor ICAD. *Nature* **391**, 43–50 (1998).
- Enari, M., Talianian, R. V., Wong, W. W. & Nagata, S. Sequential activation of ICE-like and CPP32-like proteases during Fas-mediated apoptosis. *Nature* **380**, 723–726 (1996).
- Longthorne, V. & Williams, G. Caspase activity is required for commitment to Fas-mediated apoptosis. *EMBO J.* **16**, 3805–3812 (1997).
- Armstrong, R. C. *et al.* Fas-induced activation of the cell death-related protease CPP32 is inhibited by Bcl-2 and by ICE family protease inhibitors. *J. Biol. Chem.* **271**, 16850–16855 (1996).
- Thornberry, N. A. *et al.* A combinatorial approach defines specificities of members of the caspase family and granzyme B. *J. Biol. Chem.* **272**, 17907–17911 (1997).
- Talianian, R. V. *et al.* Substrate specificities of caspase family proteases. *J. Biol. Chem.* **272**, 9677–9682 (1997).

- Chinnaiyan, A. M. *et al.* Molecular ordering of the cell death pathway. *J. Biol. Chem.* **271**, 4573–4576 (1996).
- Koopman, G. *et al.* Annexin V for flow cytometric detection of phosphatidylserine expression on B cells undergoing apoptosis. *Blood* **84**, 1415–1420 (1994).
- Mosmann, T. Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *J. Immunol. Methods* **65**, 55–63 (1983).
- Liu, X., Zou, H., Slaughter, C. & Wang, X. DFF, a heterodimeric protein that functions downstream of caspase-3 to trigger DNA fragmentation during apoptosis. *Cell* **89**, 175–184 (1997).
- Ellis, R. E., Yuan, J. & Horvitz, H. R. Mechanisms and functions of cell death. *Annu. Rev. Cell Biol.* **7**, 663–698 (1991).
- Ucker, D. S. *et al.* Genome digestion is a dispensable consequence of physiological cell death mediated by cytotoxic T lymphocytes. *Mol. Cell Biol.* **12**, 3060–3069 (1992).
- Peitsch, M., Mannherz, H. & Tschoopp, J. The apoptosis endonucleases: cleaning up after cell death? *Trends Cell Biol.* **4**, 37–41 (1994).
- Cohen, J. J., Duke, R. C., Fadok, V. A. & Sellins, K. S. Apoptosis and programmed cell death in immunity. *Annu. Rev. Immunol.* **10**, 267–293 (1992).
- Higuchi, R. in *PCR Protocols: A guide to Methods and Applications* 177–188 (Academic, San Diego, 1990).
- Blancar, M. A. & Rutter, W. J. Interaction cloning: identification of a helix-loop-helix zipper protein that interacts with c-Fos. *Science* **256**, 1014–1018 (1992).
- Mizushima, S. & Nagata, S. pEF-BOS: a powerful mammalian expression vector. *Nucleic Acids Res.* **18**, 5322 (1990).
- Itoh, N. *et al.* The polypeptide encoded by the cDNA for human cell surface antigen Fas can mediate apoptosis. *Cell* **66**, 233–243 (1991).
- Lamarre, D. *et al.* Structural and functional analysis of poly(ADP-ribose) polymerase: an immunological study. *Biochim. Biophys. Acta* **950**, 147–160 (1988).

Acknowledgements. We thank R. V. Talianian for the caspase 3 expression plasmid, G. G. Poirier for anti-human poly(ADP-ribose) polymerase, M. A. Blancar for pGEX-2T[128/129], and S. Kumagai for secretarial assistance. This work was supported in parts by Grants-in-Aid from the Ministry of Education, Science, Sports and Culture in Japan.

Correspondence: all requests for materials should be addressed to S.N. (e-mail: nagata@genetic.med.osaka-u.ac.jp).

Poly(A)- and poly(U)-specific RNA 3' tail shortening by *E. coli* ribonuclease E

Hongjin Huang*, Jian Liao† & Stanley N. Cohen*†

Departments of * Genetics and † Medicine, Stanford University School of Medicine, Stanford, California 94305-5120, USA

Ribonuclease (RNase) E is an extensively studied enzyme from *Escherichia coli* whose site-specific endoribonuclease activity on single-stranded RNA has a central role in the processing of ribosomal RNA, the degradation of messenger RNA and the control of replication of ColE1-type plasmids (for recent reviews, see refs 1–3). Here we report a previously undetected activity of RNase E: the ability to shorten 3' poly(A)- and poly(U)-homopolymer tails on RNA molecules. This activity, which leaves a 6-nucleotide adenylate or a 1-nucleotide uridylate remnant on primary transcripts, resides in the amino-terminal region of RNase E and does not require other protein cofactors. Addition of a 3'-terminal phosphate group prevents both removal of the poly(A) tail and endonucleolytic cleavage within primary transcripts, but has no effect on the cleavage of transcripts with tails that have already been truncated. The ability of RNase E to shorten poly(A) tails, together with the effect of tail length on endonucleolytic cleavage within primary transcripts, suggests a mechanism by which RNase E may exercise overall control over RNA decay.

During our investigation of RNase E cleavage of 3' polyadenylated RNAI, the antisense repressor of replication of ColE1-type plasmids^{4,5}, we observed that poly(A) tails were dramatically shortened under assay conditions in which polynucleotide phosphorylase (PNPase) and RNase II—the two *E. coli* enzymes that have 3'-to-5' exonucleolytic activity^{1,2,6–9}—were not expected to function. To investigate this observation, histidine-tagged full-length RNase E and its N-terminal endoribonucleolytic domain (N-Rne) were overexpressed in *E. coli* and affinity-purified. The N-Rne polypeptide, which contains the first 498 amino acids of the 1,061-amino-acid protein¹⁰ but lacks the PNPase binding site at the C-terminal end of RNase E¹¹, was further purified to homogeneity by elution from preparative SDS–polyacrylamide gels, followed by renaturation

(Fig. 1a); examination of gel-purified preparations by two-dimensional gel electrophoresis (Fig. 1b) and western immunoblotting (Fig. 1c) revealed only RNase E isoforms. Additional western immunoblotting using anti-PNPase or anti-RNase II antibodies failed to detect either of these enzymes.

Addition of gel-purified N-Rne to a 111-nucleotide 5'-³²P-labelled RNAI derivative (GGG.RNAI (ref. 12); Fig. 2a) containing a 3' tail consisting of 40 or 20 adenylates produced digestion ladders showing time-dependent cleavage of a series of phosphodiester bonds a single nucleotide apart (Fig. 2b, c); furthermore, digestion of poly(A) tails occurred in reaction mixtures lacking phosphate—which is required for PNPase activity—and K⁺, which is required for RNase II activity (refs 6, 7; and our data not shown). Shortening of poly(A) tails of different lengths stopped 6 nucleotides from the base of the hairpin loop structure which is present at the natural terminus of RNAI (Fig. 2b, c); poly(A) tails of this length predominate on RNAI molecules isolated from *E. coli* cells^{13,14}. The detection of a full series of decay intermediates with tails of 6–20 adenylates (Fig. 2c) and the paucity of decay intermediates containing more than 20 adenosines even at short digestion times, together suggest that removal of the distal adenylates of poly(A) tails may occur more rapidly than removal of the proximal ones.

Notwithstanding the shortening of the single-strand stretch of adenylates at the 3' end of GGG.RNAI, the 8-nucleotide single-strand fragment GGGACAGU, cleaved from the 5' end by the endoribonucleolytic activity^{5,14} of gel-purified N-Rne, was not degraded (Fig. 2b); this attests to the specificity of the attack on the poly(A) tail and the absence of PNPase or RNase II activity, both of which digest single-stranded RNA segments nonspecifically^{6–8} and can degrade the 8-nucleotide fragment (our unpublished results). This fragment did not decay in these reaction mixtures even when phosphate or K⁺ were subsequently added, confirming the absence of RNase II and PNPase.

GGG.RNAI was polyadenylated using [α -³²P]ATP and poly(A) polymerase I (PAP I), and poly(A)-tailed GGG.RNAI molecules (~150 nucleotides long) isolated from urea/acrylamide gels were digested with gel-purified N-Rne (Fig. 2d). The release of products consisting of more than 70% mononucleotides (on a molar basis, by densitometric analysis) confirmed the ability of RNase E to carry out multiple cleavages at adjacent phosphodiester bonds between adenylate residues in the homopolymer tail, as commonly observed during exonucleolytic digestion. Additionally, this digestion generated a ladder of radioactive low-molecular-weight oligomers, indicating that the enzyme can also attack internal phosphodiester bonds in poly(A) tails.

A 3'-terminal addition of 40 uridines to GGG.RNAI was degraded by cleavage of adjacent phosphodiester bonds faster and more completely than a poly(A) tail of the same length, yielding molecules containing only one unpaired U at the 3' end (Fig. 2b, c). There was no digestion of poly(G) or poly(C) tails by RNase E under the same assay conditions (Fig. 2b), indicating that homopolymer removal for the 3' ends of RNA molecules is sequence-specific. However, a 3' poly(G) addition to GGG.RNAI unmasked a cryptic site of RNase E-mediated endonucleolytic cleavage in the primary transcript.

The ability of N-Rne to remove poly(A) tails and its endonucleolytic activity are congruent with the amount of N-Rne protein in elution fractions (Fig. 1a, and data not shown). Replacement of Mg²⁺ with Ca²⁺ resulted in the loss of both activities. Mutational truncation of N-Rne by removal of the region between amino acids 321 and 498, which eliminates the endoribonucleolytic activity of RNase E (ref. 10), produced a product, SN-Rne; this product, when overexpressed and either affinity-purified or gel-purified in the same way as N-Rne (Fig. 1a), was also unable to remove poly(A) tails. We have thus identified the RNase E sequences required for poly(A) tail removal and shown that this activity lies in the same region of the protein as its endonucleolytic activity. We

have not been able to separate the poly(A)-tail shortening and endonucleolytic activities of N-Rne; however, ARD-1, a human RNA-binding protein that cleaves a variety of substrates, including RNAI, at the same phosphodiester bonds as does RNase E¹⁵, does not shorten poly(A) tails (data not shown). Therefore, enzymatic ability to cleave at the same sites as RNase E is not congruent *per se* with an ability to shorten poly(A) tails.

We did not detect any radioactively labelled mononucleotides or 3'-labelled fragments generated by N-Rne digestion of GGG.RNAI-40A ligated at its 3' terminus with either 5'-[³²P]Ap-3' (Fig. 3) or 5'-[³²P]Cp-3'. Similarly, no mononucleotides were generated by N-Rne digestion of GGG.RNAI-6A ligated at its 3' terminus with 5'-[³²P]Ap-3'. However, [³²P]Ap-3'-terminated GGG.RNAI-6A, but not [³²P]Ap-3'-terminated GGG.RNAI-40A, was attacked endonucleolytically by RNase E at a previously identified cleavage site in the primary transcript^{5,14,16}, generating a ³²P-labelled 3' fragment of the expected length (Fig. 3). The failure of RNase E digestion of RNAI-40A-[³²P]Ap-3' to produce any detectable 3'-labelled decay products (compare with the decay intermediates of 5'-[³²P]RNAI-40A-OH-3') implies that the presence of a 3' phosphate group precludes both poly(A) tail shortening and endonucleolytic digestion of the primary transcript by RNase E. Taken together, these findings suggest that the length of the poly(A) tail on RNAI derivatives can modulate endonucleolytic attack on primary transcripts by RNase E, and consequently that poly(A)-tail shortening may be required for degradation of the primary transcript, as in eukaryotic RNA decay¹⁷.

It has been proposed that *E. coli* has a 3' ribonuclease specifically for shortening poly(A) tails^{13,18,19}, but our discovery that RNase E—which has been studied for more than 20 years—has this activity was unexpected. Completion of poly(A) tail removal and 3' to 5' degradation of primary transcripts requires a separate enzyme, which on the basis of previous results^{13,18,20,21} may be PNPase, RNase E's companion in a multicomponent 'degradosome'^{22–25} complex. As 3'-adenylation facilitates PNPase attack on RNAI¹³, the 6-nucleotide poly(A) remnant left by the tail-shortening activity of RNase E may aid 3'-to-5' PNPase digestion of primary transcripts. The release by RNase E of short oligomers containing 2–10

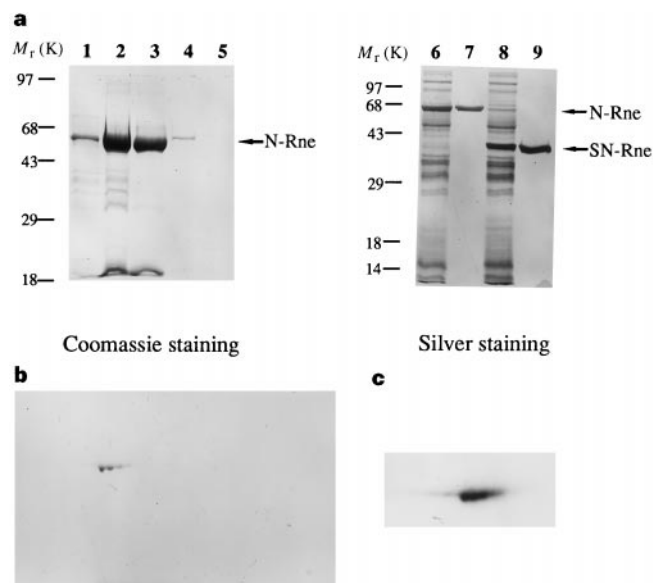
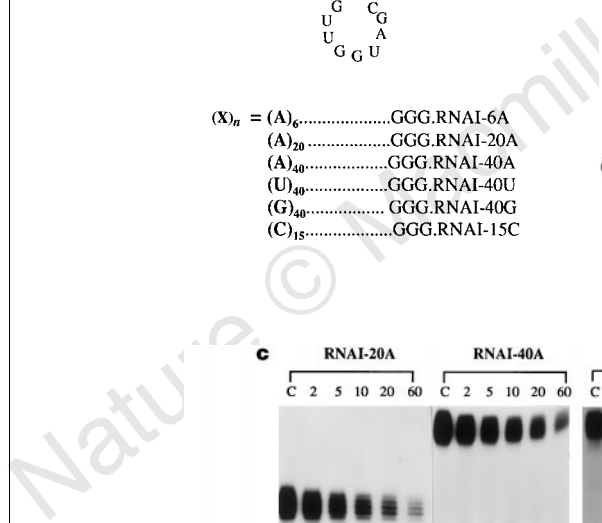


Figure 1 Purification of RNase E. **a**, Lanes 1–5, SDS-PAGE analysis of the elution fractions from affinity chromatography (Coomassie blue staining); lanes 6 and 8, whole-cell lysates of *E. coli* BL21(DE3) overexpressing polypeptide containing the first 498 (N-Rne) or 321 (SN-Rne) amino acids by RNase E, respectively; lanes 7 and 9, the corresponding proteins purified from SDS-polyacrylamide gels (lanes 6–9, silver staining). **b**, Two-dimensional gel analysis, and **c**, 2-dimensional immunoblot analysis using anti-RNase E mAb 116.7 to verify the purity of the protein.

The separate ends of RNAI, and perhaps of other prokaryotic RNA molecules as well, may normally be held in close proximity, possibly by the degradosome^{13,14,18,25}. If the termini of polyadenylated prokaryotic mRNAs are in fact linked in *E. coli* cells, as has been suggested for at least some eukaryotic mRNAs²⁷, the shortening of poly(A) tails by RNase E may cause separation of the RNA ends, permitting endonucleolytic cleavage by RNase E and also facilitating 3'-to-5' degradation of both the truncated poly(A) tail and the primary transcript by PNPase. Our finding that RNAI transcripts containing long poly(A) additions are not cleaved endonucleolyti-



associated with poly(G) addition. **c**, Extended electrophoresis of the samples RNAI-20A, RNAI-40A and RNAI-40U digested by RNase E for the times indicated (in minutes) to show a ladder of decay intermediates differing in length by 1 nucleotide; arrows indicate migration positions of RNAI-6A and RNAI. **d**, N-Rne digestion of RNAI labelled at the 3' terminus by [α - 32 P]ATP using poly(A) polymerase I. Digestion times are shown across the top (in minutes). Arrow (right) indicates the position of mononucleotides released from the 3' terminus. Left, PNase digestion of continuously labelled RNAI-40A; the position of mononucleoside diphosphates is indicated.

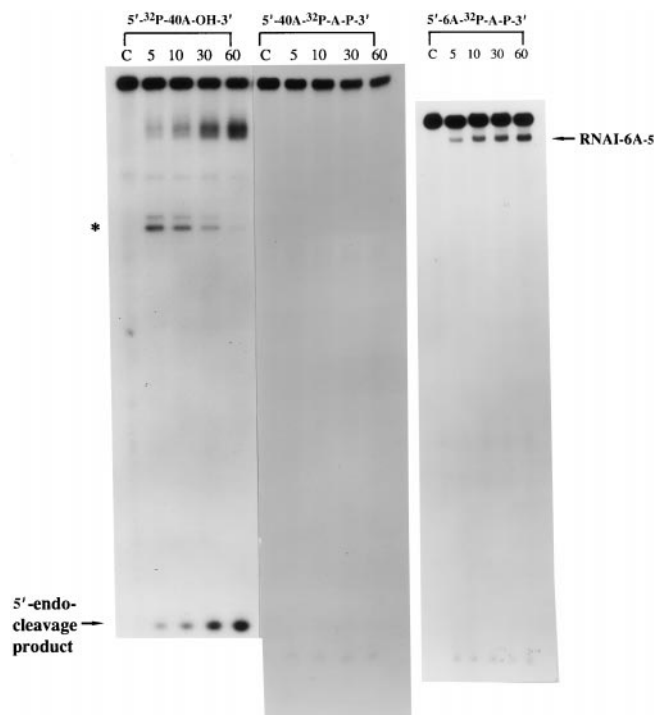


Figure 3 Elimination of both poly(A)-tail removal and endonucleolytic activity of N-Rne by 3' terminal addition of [32 P]A_p to RNAI-40A. Asterisk indicates bands produced by internal cleavage at minor sites. Digestion times are indicated in minutes. For comparison, RNAI-6A- 32 P]A_p was cleaved at its 5' end by RNase E, as previously identified, and the 3' product is indicated (right arrow). The result was the same for RNAI-Pap (data not shown).

cally by RNase E, whereas identical molecules having short tails or no tails are, is consistent with this notion. Thus, the shortening of poly(A) tails by RNase E and its endonucleolytic cleavage of the primary transcript may together exercise overall control over the rate of RNA decay in *E. coli*. □

Methods

Purification and analysis of RNase E. *E. coli* BL21 (DE3) cells (Novagen) containing pET-16b based plasmids encoding the first 498 and 321 amino acids were used for purification of the N-Rne and SN-Rne polypeptides, respectively¹⁰. The procedure used non-denaturing immobilized metal affinity chromatography (IMAC)²⁸, as described in the pET System Manual (Novagen). His-tagged N-Rne was eluted primarily at 0.2–0.3 M of imidazole. Fractions were individually dialysed against 50 mM Tris-HCl, pH 8.0, 0.1 mM EDTA, 0.15 M NaCl, 1 mM DTT and 20% glycerol. Fraction 3 in Fig. 1a containing 90% pure N-Rne protein was used for the gel purification procedure²¹. The protein was renatured by dialysing against the same buffer but containing decreasing concentrations of denaturant. Two-dimensional gel analysis was done using a gel apparatus (Mini-PROTEAN II; BioRad) by following the manufacturer's protocol; anti-RNase E mAb 116.7 was diluted 2,000 fold for western immunoblotting.

Synthesis of substrates. DNA templates used for *in vitro* production (Amicon T7 MEGA Kit) of RNAI, RNAI-6A, RNAI-40A, RNAI-40U, RNAI-40G and RNAI-15C were PCR fragments based on plasmid pM21 (ref. 12). The T7 sequence TAATACGACTCATAATAGGGACAGTATTGGTATCTGC was used as the forward primer; reverse primers were AACAAAAAACCACCGCTAC, (T)₆AACAAAAAACCACCGCTAC, (T)₄₀AACAAAAAACCACCGCTAC, (A)₄₀AACAAAAAACCACCGCTAC, (C)₄₀AACAAAAAACCACCGCTAC, and (G)₁₅AACAAAAAACCACCGCTAC, respectively. For 5' labelling, RNA transcribed *in vitro* was 5'-dephosphorylated using calf intestinal alkaline phosphatase (Life Technologies) then labelled with [γ - 32 P]ATP using T4 polynucleotide kinase, and purified from 8% denaturing polyacrylamide gels. For 3'-end labelling, 5'-[32 P]A_p-3' was ligated to unlabelled RNAI-40A

using T4 RNA ligase (Boehringer-Mannheim). Poly(A)polymerase I (PAP I) was also used to label RNAI at the 3' end, as follows: histidine-tagged PAP I was purified by non-denaturing IMAC²⁸ and dialysed against 25 mM Tris-HCl (pH 8.0), 0.5 M NaCl, 1 mM EDTA, 0.1 mM DTT and 50% glycerol. 150 pmol unlabelled RNAI was incubated with 50 ng PAP I and 50 μ Ci [γ - 32 P]ATP in 50 μ l reaction buffer consisting of 50 mM Tris (pH 8.0), 10 mM MgCl₂, 2.5 mM MnCl₂, 0.25 M NaCl and 20 μ g tRNA; reactions were stopped by addition of phenol at 5, 10, 20 and 40 min, respectively. Polyadenylated RNAI molecules of ~150 nt long (containing ~40 32 P-labelled adenosines) were purified from gels.

Enzymatic assays. Labelled substrate (5 pmol) was incubated at 30 °C with 15 ng of renatured, gel-purified N-Rne in reaction buffer consisting of 20 mM Tris-HCl (pH 8.0), 10 mM MgCl₂, 100 mM NaCl, 0.1 mM DTT and 80 μ g ml⁻¹ yeast tRNA. Aliquots (8 μ l) were removed from each reaction, mixed separately with 8 μ l sequencing 'stop' buffer, denatured by heating at 85 °C for 3 min, and separated on 9% urea/polyacrylamide gels (run at 50 W for 2 h). To detect RNA species differing in length by 1 nt, the electrophoresis was run for longer (at 50 W for 5 h). PNPase was assayed in 10 μ l of reaction buffer containing 20 mM Tris-HCl (pH 7.5), 1 mM MgCl₂, 20 mM KCl, 10 mM K₂HPO₄, 1 mM ATP and 1 mM DTT, plus 1 μ l PNPase (Sigma) diluted 300-fold.

Received 18 August; accepted 29 September 1997.

- Cohen, S. N. & McDowell, K. J. RNase E: still a wonderfully mysterious enzyme. *Mol. Microbiol.* **23**, 1099–1106 (1996).
- Ehretsmann, C. P., Carpousis, A. J. & Krusch, H. M. mRNA degradation in prokaryotes. *FASEB J.* **6**, 3186–3192 (1992).
- Meleforts, O., Lundberg, U. & von Gabain, A. in *Control of Messenger RNA Stability* (eds Belasco, J. & Brawerman, G.) 53–70 (Academic, San Diego, 1993).
- Tomcsanyi, T. & Apirion, D. Processing enzyme ribonuclease E specifically cleaves RNA I: an inhibitor of primer formation in plasmid DNA synthesis. *J. Mol. Biol.* **185**, 713–720 (1985).
- Lin-Chao, S. & Cohen, S. N. The rate of processing and degradation of antisense RNAI regulates the replication of ColE1 type plasmids *in vivo*. *Cell* **65**, 1233–1242 (1991).
- Littauer, U. Z. & Soreq, H. Polynucleotide phosphorylase. *The Enzymes* **XV**, 517–553 (1982).
- Singer, M. F. & Tolbert, G. Purification and properties of a potassium-activated phosphodiesterase (RNase II) from *Escherichia coli*. *Biochemistry* **4**, 1319–1330 (1965).
- Nossal, N. G. & Singer, M. F. The processive degradation of individual polyribonucleotide chains. *J. Biol. Chem.* **243**, 913–922 (1968).
- Klee, C. B. & Singer, M. F. The processive degradation of individual polyribonucleotide chains. *J. Biol. Chem.* **243**, 923–927 (1968).
- McDowell, K. J. & Cohen, S. N. The N-terminal domain of the *rne* gene product has RNase E activity and is non-overlapping with the arginine-rich RNA-binding motif. *J. Mol. Biol.* **255**, 349–355 (1996).
- Kido, M., Yamanaka, K., Mitani, T., Niki, H., Ogura, T. & Hiraga, S. RNase E polypeptides lacking a carboxyl-terminal half suppress a *mukB* mutation in *Escherichia coli*. *J. Bacteriol.* **178**, 3917–3925 (1996).
- Helmer-Citterich, M., Anceschi, M. M., Banner, D. W. & Cesareni, G. Control of ColE1 replication: low affinity specific binding of Rop (Rom) to RNAI and RNAII. *EMBO J.* **7**, 557–566 (1988).
- Xu, F. & Cohen, S. N. RNA degradation in *Escherichia coli* regulated by 3' adenylation and 5' phosphorylation. *Nature* **374**, 180–183 (1995).
- Xu, F., Lin-Chao, S. & Cohen, S. N. The *Escherichia coli* *pcnB* gene promotes adenylation of antisense RNAI of ColE1-type plasmids *in vivo* and degradation of RNAI decay intermediates. *Proc. Natl Acad. Sci. USA* **90**, 6756–6760 (1993).
- Claverie-Martin, F., Wang, M. & Cohen, S. N. Ard-1 cDNA from human cells encodes a site-specific single strand endonuclease that functionally resembles *Escherichia coli* RNase E. *J. Biol. Chem.* **272**, 13823–13828 (1997).
- Kaberdin, V. R., Chao, Y.-H. & Lin-Chao, S. RNase E cleaves at multiple sites in bubble region of RNAI stem-loops yielding products and dissociate differentially from the enzyme. *J. Biol. Chem.* **271**, 13103–13109 (1996).
- Caponigro, G. & Parker, R. Mechanisms and control of mRNA turnover in *Saccharomyces cerevisiae*. *Microbiol. Rev.* **60**, 233–249 (1996).
- Cohen, S. N. Surprises at the 3' end of prokaryotic RNA. *Cell* **80**, 829–832 (1995).
- Sarkar, N. Polyadenylation of mRNA in bacteria. *Microbiology* **142**, 3125–3133 (1996).
- Hajnsdorf, E., Braun, E., Haug-Nielsen, J. & Regnier, P. Polyadenylation destabilizes the *rpsO* mRNA of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **92**, 3973–3977 (1995).
- Coburn, G. A. & Mackie, G. A. Overexpression, purification, and properties of *Escherichia coli* ribonuclease II. *J. Biol. Chem.* **271**, 1048–1053 (1996).
- Py, B., Higgins, C. F., Krusch, H. M. & Carpousis, A. J. A DEAD-box RNA helicase in the *Escherichia coli* RNAI degradation. *Nature* **381**, 169–172 (1996).
- Carpousis, A. J., Van Houwe, G., Ehretsmann, G. & Krusch, H. M. Copurification of *E. coli* RNase E and PNPase: evidence for a specific association between two enzymes important in RNA processing and degradation. *Cell* **76**, 889–900 (1994).
- Miczak, A., Kaberdin, V. R., Wei, C. L. & Lin-Chao, S. Proteins associated with RNase E in a multicomponent ribonucleolytic complex. *Proc. Natl Acad. Sci. USA* **93**, 3865–3869 (1996).
- Py, B., Causton, H., Mudd, E. A. & Higgins, C. F. A protein complex mediating mRNA degradation in *Escherichia coli*. *Mol. Microbiol.* **14**, 717–799 (1994).
- Razzell, W. E. & Khorana, H. G. Studies on polynucleotides. III. Enzymatic degradation. Substrate specificity and properties of snake venom phosphodiesterase. *J. Biol. Chem.* **234**, 2105–2113 (1959).
- Jacobson, A. Poly(A) metabolism and translation: The closed-loop model. in *Cold Spring Harbor Monogr. Ser.* **30**, 451–480 (1996).
- Hochuli, E., Dobeli, H. & Schacher, A. New metal chelate adsorbent selective for proteins and peptides containing neighboring histidine residues. *J. Chromatogr.* **411**, 177–184 (1987).

Acknowledgements. These studies were supported by a grant from the US NIH. We thank Y. Feng for hist-tagged PAP I constructs, and J. Belasco, R. Lehman and C. Higgins for comments on this manuscript.

Correspondence and requests for materials should be addressed to S.N.C. (e-mail: sncohen@forsythe.stanford.edu).

Structure of IRF-1 with bound DNA reveals determinants of interferon regulation

Carlos R. Escalante*, Junming Yie†, Dimitris Thanos† & Aneel K. Aggarwal*

* Structural Biology Program, Department of Physiology and Biophysics, Box 1677, Mount Sinai School of Medicine, 1425 Madison Avenue, New York, New York 10029, USA

† Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, USA

The family of interferon regulatory factor (IRF) transcription factors is important in the regulation of interferons in response to infection by virus and in the regulation of interferon-inducible genes^{1,2}. The IRF family is characterized by a unique 'tryptophan cluster' DNA-binding region. Here we report the crystal structure of the IRF-1 region bound to the natural positive regulatory domain I (PRD I) DNA element from the interferon- β promoter¹. The structure provides the first three-dimensional view of a member of the growing IRF family, revealing a new helix–turn–helix motif that latches onto DNA through three of the five conserved tryptophans. The motif selects a short GAAA core sequence through an obliquely angled recognition helix, with an accompanying bending of the DNA axis in the direction of the protein. Together, these features suggest a basis for the occurrence of GAAA repeats within IRF response elements and provide clues to the assembly of the higher-order interferon- β enhancosome.

IRF-1 functions as a transcriptional activator when bound to the interferon- β (IFN- β) promoter. IRF-1 recognizes two regulatory elements (PRD I and PRD III) and synergizes with the activators ATF-2/Jun and NF- κ B, which bind to the nearby PRD IV and PRD II elements, respectively (Fig. 1a)¹. These activators, together with the high-mobility group protein HMG I (Y), assemble to form a higher-order nucleoprotein complex, the enhancosome, leading to the specific activation of the IFN- β gene³. IRF-1 is also the founder member of the IRF family of transcription factors, which now includes IRF-2, the IFN consensus sequence binding protein

(ICSBP), the IFN- α stimulated gene factor 3 γ (ISGF3 γ), IRF-3, the PU.1 interaction partner (Pip), IRF-5 and IRF-7 (refs 4–9). The IRF family contains a homologous DNA binding region characterized by a signature of five tryptophans or a 'tryptophan cluster' (Fig. 1b); otherwise its sequence is unrelated to other transcription factor classes. We cocrystallized IRF-1 (residues 1–113) with a 13-base pair (bp) DNA fragment containing the natural PRD I sequence taken from the IFN- β promoter¹ (Fig. 1a). The structure, determined at 3 Å resolution, shows IRF-1 bound to the PRD I element as a monomer with the asymmetric unit containing two such complexes related by a pseudo-dyad.

The IRF-1 DNA-binding region has an α/β architecture comprising a cluster of three α -helices (α 1– α 3) flanked on one side by a mixed four-stranded β -sheet (β 1– β 4) (Fig. 2). Topologically, the fold is similar to the helix–turn–helix (HTH)-containing DNA-binding domains of catabolite gene activator protein (CAP)-related proteins such as histone H5 and the hepatocyte nuclear factor 3 γ (HNF-3 γ)^{10–12}. Despite the topological similarity, the shape of the IRF-1 fold and its mode of DNA interaction are distinct from the other CAP-related proteins. Most striking is the presence of three large loops (L1, L2 and L3) that emanate between β 2 and α 2, α 2 and α 3, and β 3 and β 4, respectively (Fig. 2). Loop L3 has a counterpart in HNF-3 γ (Fig. 2c), but loops L1 and L2 appear to be characteristic traits of the IRF family of transcription factors (Fig. 1b). Helices α 2 and α 3 comprise the first and second helices of the HTH motif, with α 3 (also known as the recognition helix) lying in the DNA major groove. The orientation of α 3 differs radically from that observed in canonical HTH proteins. In most HTH proteins, the recognition helix traverses the major groove with an orientation perpendicular to the DNA axis¹³. In contrast, the IRF-1 recognition helix tracks the major groove in a tangential fashion with its axis almost parallel to the DNA sugar-phosphate backbone (Fig. 2). This 'angle of attack' is reminiscent of the way the basic α -helices of bHLH transcription factor Max follow the major groove¹⁴.

Contacts to bases within the major groove are localized to a GAAA core sequence within the 13-bp PRD I element (Fig. 3). The amino acids mediating these contacts in the major groove (Arg 82, Cys 83, Asn 86 and Ser 87) project from the last two turns of the IRF-1 recognition helix. Arg 82 is in a position to form bidentate hydrogen bonds with the guanine of the outer G:C base pair (GAAA), while establishing a salt link with the preceding 5'

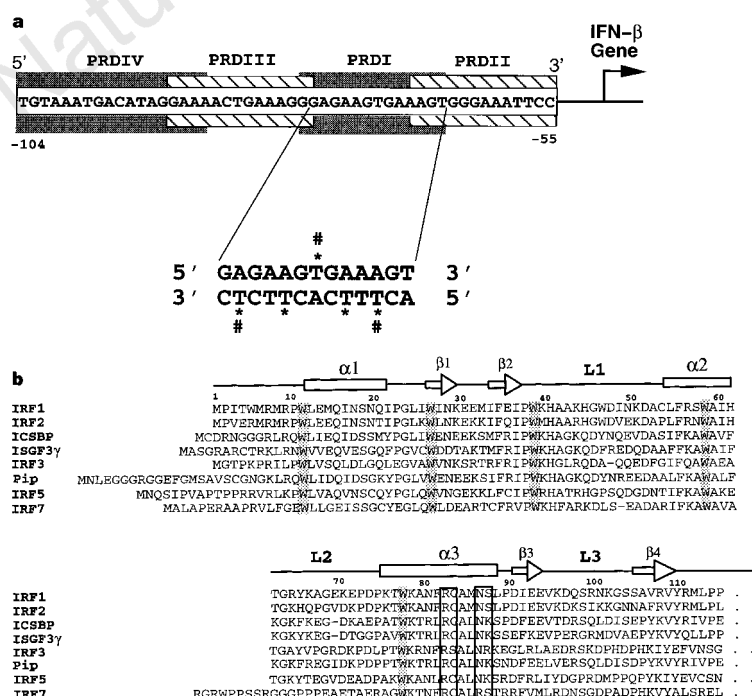


Figure 1 a, At the top is shown the sequence of the IFN- β promoter containing the regulatory elements PRD I-IV (ref. 1). At the bottom is shown the sequence of the 13-bp oligomer used in co-crystallization studies. Asterisks and daggers denote the thymine residues that were substituted by bromouracils and iodouracils, respectively (see Table 1). **b**, The amino-acid sequence of IRF-1 DNA-binding region (residues 1–113) aligned with that of other members of the IRF family^{4–9}. The sequences of IRF-5 and IRF-7 were obtained from the Swiss-protein database (accession numbers Q13568 and Q92985, respectively). Shown above the sequences are the relative locations of α -helices, β -strands and loops in IRF-1. The structure verifies the topology predicted for IRF-2 by NMR³⁰. The tryptophans that comprise the 'tryptophan cluster' of the IRF family have been shaded. The residues contacting DNA bases in the major groove are emphasized in open boxes.

phosphate group (Fig. 3a). Recognition of the second and third base pairs of the core sequence is due primarily to Cys 83, which through its sulphhydryl group receives a single hydrogen bond from N6 of adenine at the second position (GAAA) and donates a hydrogen bond to O4 of thymine at the third position (GAAA). Arg 82 is totally conserved within the IRF family whereas Cys 83 shows only a single substitution to serine in IRF-3 (Fig. 1b), consistent with the 'promiscuous' DNA binding observed among IRF family members^{5,15}. A mutation of either residue to alanine is found to abrogate DNA binding (J.Y., C.R.E., A.K.A. and D.T., manuscript in preparation). Specificity for an A:T base pair at the fourth position (GAAA) is conferred by van der Waals contacts between the methyl group of the thymine and the O γ atom of Ser 87 (Fig. 3a). In some IRF family members, Ser 87 is substituted by Lys or Arg, which has the potential to reach over and mediate hydrogen bonding outside the GAAA core sequence. Supplementing the direct contacts in the major groove, a water-mediated interaction between Asn 86 and the N7 of adenine at the second position (GAAA) is observed in both IRF-1–DNA complexes within the asymmetric unit. There is also the possibility of minor groove contacts due to His 40 from loop L1 reaching into the minor groove adjacent to the major groove in which the recognition helix is located (Fig. 3a). However, the distances between His 40 and the DNA bases are long (>3.8 Å), reminiscent of the DNA-binding domains of phage 434 repressor and tumour suppressor p53s in which solitary arginine residues

penetrate the minor groove but not far enough to make direct contacts with the bases^{16,17}. Taken together, the structure of the IRF-1–DNA complex provides a basis for GAAA sequences occurring within almost all IRF response elements. Curiously, some response elements, such as the PRD III element bound by IRF-1 and the interferon-stimulated response elements (ISREs) bound by ISGF3 γ , contain two GAAA core sequences^{1,18}, suggesting that these elements may accommodate two IRF molecules.

Calculation of the change in solvent-accessible surface area upon DNA binding gives a value of about 2,000 Å² for IRF-1. This value is greater than that calculated for monomeric homeodomain–DNA complexes (~1,500 Å²) and reflects the extensive contacts IRF-1 makes with the DNA backbone. In addition to the three loops, residues within all three α -helices form interactions with the DNA backbone (Fig. 3b). Loops L1 and L3 spread out from the body of the protein and contact phosphate groups as far as six nucleotides upstream of the GAAA core sequence. This may be the reason why IRF-1 has an extended footprint^{3,5} even though it only contacts 4 bp within the PRD I sequence. The loops presumably become ordered upon DNA binding, imposing a substantial entropy penalty that is paid through the release of water molecules and cations bound to the DNA backbone. The intrinsic flexibility of the loops is evident from their high r.m.s.d., 1.02 Å versus 0.27 Å for core C α s, when comparing the two molecules in the asymmetric unit. In common with other DNA-binding proteins, many of the residues contacting

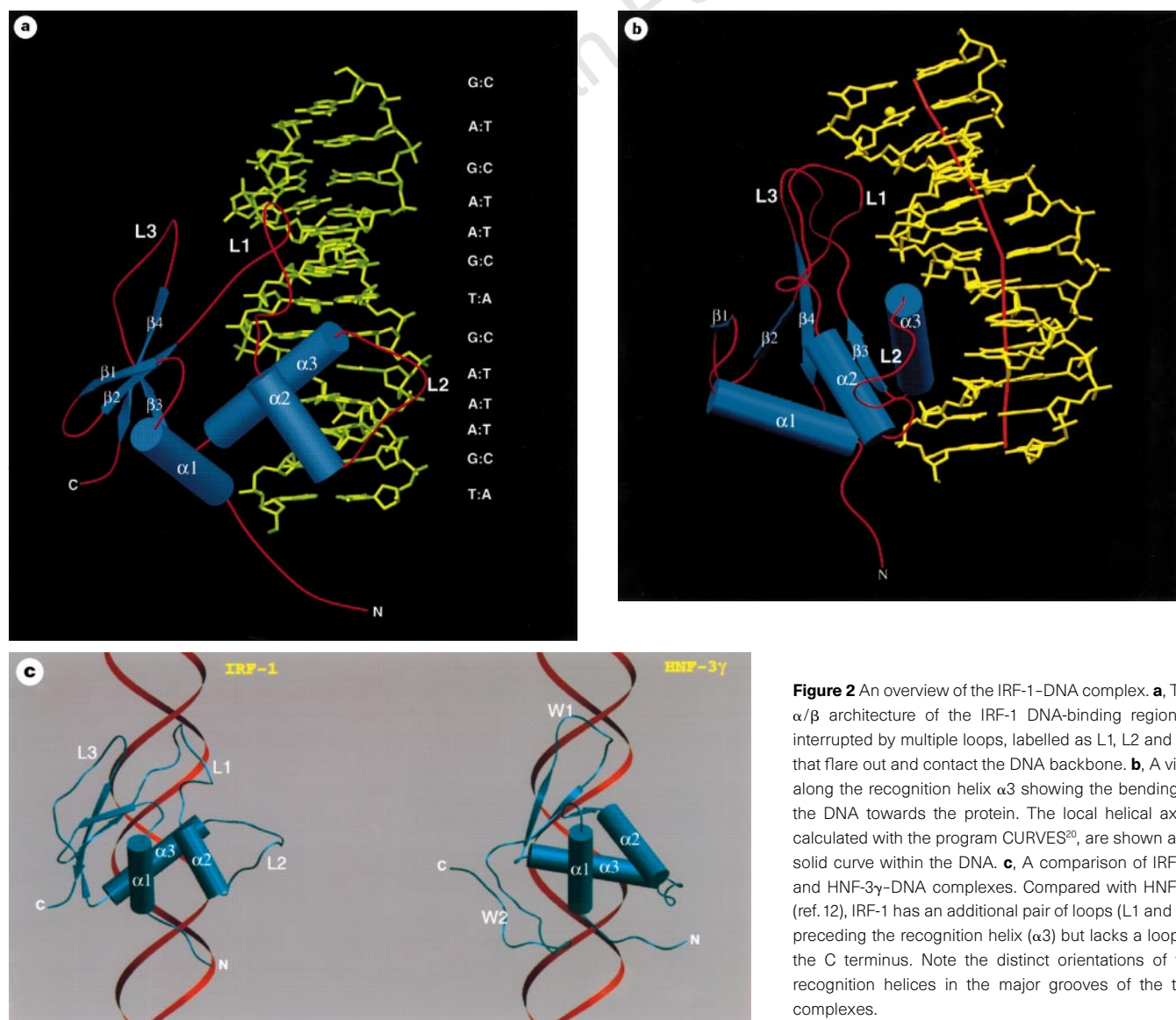
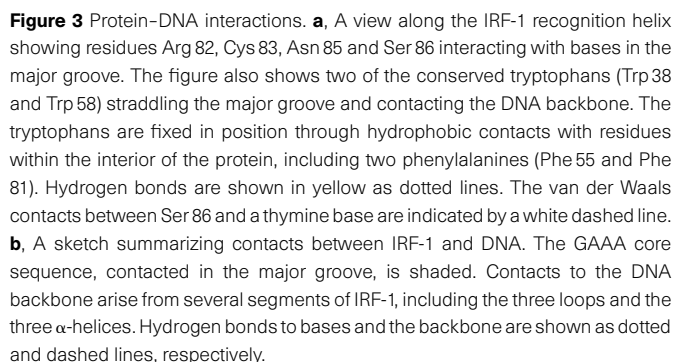


Figure 2 An overview of the IRF-1–DNA complex. **a**, The α/β architecture of the IRF-1 DNA-binding region is interrupted by multiple loops, labelled as L1, L2 and L3, that flare out and contact the DNA backbone. **b**, A view along the recognition helix α_3 showing the bending of the DNA towards the protein. The local helical axes, calculated with the program CURVES²⁰, are shown as a solid curve within the DNA. **c**, A comparison of IRF-1- and HNF-3 γ -DNA complexes. Compared with HNF-3 γ (ref. 12), IRF-1 has an additional pair of loops (L1 and L2) preceding the recognition helix (α_3) but lacks a loop at the C terminus. Note the distinct orientations of the recognition helices in the major grooves of the two complexes.



The DNA bends towards the protein by $\sim 22^\circ$ (Fig. 2b), with a narrowing of the major groove in which the recognition helix is located²⁰. The bending is greater than in the HNF-3 γ -DNA complex¹² ($\sim 13^\circ$), and has implications for how the IRF family members interact with co-transcriptional activators. For instance, the bending induced by IRF-1 at PRD I and PRD III elements will not only influence its interactions with the neighbouring NF- κ B and ATF-2/Jun molecules, but also help to juxtapose ATF-2/Jun and NF- κ B closer together within the IFN- β enhanceosome^{3,21}. Interestingly, the structure of IRF-1/PRD I complex modelled alongside that of NF- κ B p50 homodimer^{22,23} bound to PRD II reveals steric clashes between the first ten residues of IRF-1 and the dimerization domain of NF- κ B (Fig. 4). The clashes are aggravated by the cumulative bend in the DNAs of the two complexes (Fig. 4). This was borne out *in vitro*, where the two proteins were found to inhibit each other's DNA binding³ (J.Y., C.R.E., A.K.A. and D.T., manuscript in preparation). Remarkably, the interference between IRF-1 and NF- κ B could be relieved by the addition of HMG I (Y) which specifically interacts with both IRF-1 and NF- κ B³. DNA bending may also play a functional role in the interaction between ISGF3 β

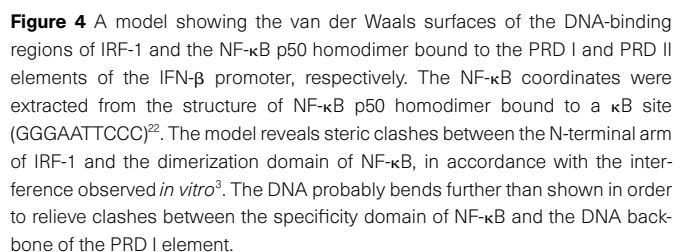


Table 1 Crystallographic analysis

Item	Br-edge	Br-peak	Br-remote	Iodo
Wavelength (Å)	0.92048	0.91980	0.90496	1.542
Resolution (Å)	20-3.0	20-3.0	20-3.0	20-3.5
Number of reflections				
Measured	125,805	125,923	127,213	117,593
Unique	21,657	21,657	21,674	7,225
Data coverage (%)				
Overall	99.1	99.1	99.2	98.1
Outer Shell (3.23-3.0 Å)	99.7	99.8	99.7	97.8
R_{sym}^*				
Overall	0.061	0.062	0.069	0.068
Outer Shell (3.23-3.0 Å)	0.090	0.092	0.104	0.082
MAD phasing statistics:				
Phasing power		1.72	2.01	
(isomorphous to 3.5 Å)†				
Phasing power	1.86	2.00	1.60	1.52
(isomorphous to 3.5 Å)‡				
Refinement statistics:				
(12 to 3 Å)				
Reflections, $F > 2\sigma(F)$	20,328			
R-factor (%)	24.2			
R-free (%)§	30.9			
Non-hydrogen atoms	2,779			
r.m.s. bond lengths (Å)	0.02			
r.m.s. bond angles (°)	2.3			

* $R_{\text{sym}} = \sum |I_{\text{obs}} - \langle I \rangle| / \sum I$, calculated for all data.

† Phasing power (isomorphous) = $\langle |F_h| \rangle / \text{r.m.s.}(\epsilon)$, where $\langle |F_h| \rangle$ is the mean calculated amplitude for the heavy-atom model and r.m.s.(ϵ) is the weighted root mean square lack of closure error.

‡ Phasing power (anomalous) = $\langle |\Delta F_h| \rangle / \text{r.m.s.}(\epsilon')$, where $\langle |\Delta F_h| \rangle$ is the mean calculated Bijvoet difference from the heavy-atom model and r.m.s.(ϵ') is the weighted root mean square lack of closure for anomalous difference.

§ Calculated with the 9% of the reflection data excluded from refinement.

(or p48) and the Stat 1:2 heterodimer in the JAK-STAT signalling pathway¹⁸. The proteins bind to the interferon-stimulated response element within a few base pairs of each other to form the ISGF3 complex that is much more stable on the DNA²⁴. In the past decade, much has been learned about the regulation of the interferon system and the role of the IRF family members. The structure of IRF-1 provides a new basis for mutational and genetic studies aimed at exploring their interactions with DNA and with other classes of transcriptional activators. □

Methods

Structure determination. IRF-1 (residues 1-113) was expressed, purified and co-crystallized with a 13-bp PRD I DNA site in 18-22% PEG 8K solution at 4°C (ref. 25). The co-crystals diffract to 3.0 Å and belong to space group R3 with unit cell dimensions of $a = 84.8$ Å, $b = 84.8$ Å, $c = 203.7$ Å. For multi-wavelength anomalous diffraction (MAD) phasing²⁶, we prepared a brominated derivative of the complex, by substituting bromouracils for five thymine residues within the PRD I DNA sequence (Fig. 1). The MAD data were measured from a single frozen co-crystal (-162°C) at Cornell High Energy Synchrotron Source (CHESS). The data were recorded using the 2k × 2k charge-coupled device (CCD) detector at beamline F2. X-ray fluorescence spectra measured directly from a mounted co-crystal were found to be very similar to those recorded from a bromine foil (NaBr). Consequently, we used spectra measured from the bromine foil at regular intervals to select the monochromator settings for the different wavelengths. The data were measured at three wavelengths corresponding to the edge (0.92048 Å) and peak (0.9198 Å) of the Br K absorption profile, plus a remote point (0.90496 Å). Because of the R3 symmetry of the co-crystal, the Freidel pairs could only be recorded by the inverse beam method (ϕ and $\phi + 180^\circ$). The data were collected by the oscillation method ($\Delta\phi = 1^\circ$), with no overlap between successive frames. Inverse data were recorded after every 4-6° of rotation. By the end of the data collection the co-crystal had been rotated by 260°. We also measured data from an iodinated derivative of the complex (Fig. 1), using an RAXIS II imaging plate detector system operating on a rotating anode X-ray source. All data sets were processed using the HKL programs (HKL Research), leading to the determination of six bromine and six iodine sites which confirmed the existence of two complexes in the asymmetric unit related by a non-crystallographic dyad. For phasing, we used the program package PHASES²⁷, and a total of four data sets: Br-edge, Br-peak, Br-remote, and Iodo. The Br-edge data set was treated as 'native' in the phasing process. The Iodo data was incorporated into the phasing by first subtracting from the

'native' data the scattering from the six bromine sites by attaching negative values to the scattering factors. The final figure of merit was 0.712 at 3.5 Å resolution. The resulting electron density map showed clear densities for the DNA and many of the protein secondary structural elements.

Refinement. The map was further improved by cycles of averaging, using the non-crystallographic twofold determined from the heavy-atom positions. Using the averaged map and subsequent partial model phase combined maps, interspersed with positional, group temperature factor, and simulated annealing refinements using XPLOR²⁸, we built both IRF-1-DNA complexes using the program O (ref. 29). Residues 1-6 and 112-113 of the proteins could not be modelled because of poor density, indicating flexibility. An anisotropic correction was applied in later stages of refinement (3 Å), and the model verified through extensive simulated annealing omit maps. A total of 10 highly localized water molecules (B -values < 10 Å²) were added to the model that fit the criteria of being found at identical positions in both complexes of the asymmetric unit.

Received 8 July; accepted 17 September 1997.

- Maniatis, T. *et al.* in *Transcriptional Regulation* (eds McKnight, S. & Yamamoto, K.) 1993-1220 (Cold Spring Harbor Laboratory Press, New York, 1992).
- Taniguchi, T., Harada, H. & Lamphier, M. Regulation of the interferon system and cell growth by the IRF transcription factors. *J. Cancer Res. Clin. Oncol.* **121**, 516-520 (1995).
- Thanos, D. & Maniatis, T. Virus induction of human IFN β gene expression requires the assembly of an enhancosome. *Cell* **83**, 1091-1100 (1995).
- Miyamoto, M. *et al.* Regulated expression of a gene encoding a nuclear factor, IRF-1, that specifically binds to the IFN- β gene regulatory elements. *Cell* **54**, 903-913 (1988).
- Harada, H. *et al.* Structurally similar but functionally distinct factors, IRF-1 and IRF-2, bind to the same regulatory elements of IFN and IFN-inducible genes. *Cell* **58**, 729-739 (1989).
- Driggers, P. H. *et al.* An interferon- γ regulated protein that binds the interferon-inducible element of the major histocompatibility class I genes. *Proc. Natl Acad. Sci. USA* **87**, 3743-3747 (1990).
- Veals, S. A. *et al.* Subunit of an alpha-interferon-responsive transcription factor is related to interferon regulatory factor and Myb families of DNA binding proteins. *Mol. Cell. Biol.* **12**, 3315-3324 (1992).
- Au, W. C., Moore, P. A., Lowther, W., Juang, Y. T. & Pitha, P. M. Identification of a member of the interferon regulatory factor family that binds to the interferon-stimulated response element and activates expression of interferon-induced genes. *Proc. Natl Acad. Sci. USA* **92**, 11657-11661 (1995).
- Eisenbeis, C. F., Singh, H. & Storb, U. Pip, a novel IRF family member, is a lymphoid-specific, PU.1-dependent transcriptional activator. *Genes Dev.* **9**, 1377-1387 (1995).
- Schultz, S. C., Shields, G. C. & Steitz, T. A. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science* **253**, 1001-1007 (1991).
- Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L. & Sweet, R. M. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362**, 219-223 (1993).
- Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412-420 (1993).
- Harrison, S. C. & Aggarwal, A. K. DNA recognition by proteins with the helix-turn-helix motif. *A. Rev. Biochem.* **59**, 933-969 (1990).
- Ferre-D'Amare, A. R., Prendergast, G. C., Ziff, E. B. & Burley, S. K. Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature* **363**, 38-45 (1993).
- Kawakami, T. *et al.* Possible involvement of the transcription factor ISGF3 gamma in virus-induced expression of the IFN-beta gene. *FEBS Lett.* **358**, 225-229.
- Aggarwal, A. K., Rodgers, D. W., Drott, M., Ptashne, M. & Harrison, S. C. Recognition of a DNA operator by repressor of phage 434: a view at high resolution. *Science* **242**, 899-907 (1988).
- Cho, Y., Gorina, S., Jeffrey, P. D. & Pavletich, N. P. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* **265**, 346-355 (1994).
- Darnell, J. E. Jr, Kerr, I. M. & Stark, G. R. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signalling proteins. *Science* **264**, 1415-1421 (1994).
- Spolar, R. S. & Record, M. T. Coupling of local folding to site-specific binding of proteins to DNA. *Science* **263**, 777-784 (1994).
- Lavery, R. & Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **6**, 63-91 (1988).
- Falvo, J. V., Thanos, D. & Maniatis, T. Reversal of intrinsic DNA bends in the IFN β gene enhancer by transcription factors and the architectural protein HMG I(Y). *Cell* **83**, 1101-1111 (1995).
- Ghosh, G., Van Duyn, G., Ghosh, S. & Sigler, P. B. The structure of NF- κ B p50 homodimer bound to a κ B site. *Nature* **373**, 303-310 (1995).
- Muller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L. & Harrison, S. C. Structure of the NF- κ B p50 homodimer bound to DNA. *Nature* **373**, 311-317 (1995).
- Horvath, C. M., Stark, G. R., Kerr, I. M. & Darnell, J. E. Jr Interactions between STAT and non-STAT proteins in the interferon stimulated gene factor 3 transcription complex. *Mol. Cell. Biol.* **16**, 6957-6964 (1996).
- Escalante, C. R., Yie, J., Thanos, D. & Aggarwal, A. K. Expressions, purification, and co-crystallization of IRF-1 bound to the interferon- β element PRD I. *FEBS Lett.* **414**, 219-220 (1997).
- Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51-58 (1991).
- Furey, W. *Phases: A Program Package for the Processing and Analysis of Diffraction Data for Macromolecules* (VA Medical Center, Pittsburgh, PA, 1993).
- Brunger, A. T. *X-PLOR, Version 3.1, A System for X-ray Crystallography and NMR* (Yale Univ. Press, New Haven, 1992).
- Jones, A. T., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110-119 (1991).
- Uegaki, K., Shirakawa, M., Harada, H., Taniguchi, T. & Kyogoku, Y. Secondary structure and folding topology of the DNA binding domain of interferon regulatory factor 2, as revealed by NMR spectroscopy. *FEBS Lett.* **359**, 184-188 (1995).

Acknowledgements. We thank the staff at CHESS for facilitating data collection, E. Jacobson, H. Viadiu and D. Wah for help with data collection, G. Mogilnitskiy for technical assistance, L. Shapiro and H. Weinstein for comments on the manuscript. Coordinates have been deposited in the Brookhaven Protein Database. Accession number 1IFI.

Correspondence and requests for material should be addressed to A.K.A. (e-mail: aggarwal@inka.mssm.edu).